CrossMark

# Phylogenetic mixtures and linear invariants for equal input models

**Marta Casanellas[1] · Mike Steel[2]**

**Abstract** The reconstruction of phylogenetic trees from molecular sequence data relies on modelling site substitutions by a Markov process, or a mixture of such processes. In general, allowing mixed processes can result in different tree topologies becoming indistinguishable from the data, even for infinitely long sequences. However, when the underlying Markov process supports linear phylogenetic invariants, then provided these are sufficiently informative, the identifiability of the tree topology can be restored. In this paper, we investigate a class of processes that support linear invariants once the stationary distribution is fixed, the 'equal input model'. This model generalizes the 'Felsenstein 1981' model (and thereby the Jukes–Cantor model) from four states to an arbitrary number of states (finite or infinite), and it can also be described by a 'random cluster' process. We describe the structure and dimension of the vector spaces of phylogenetic mixtures and of linear invariants for any fixed phylogenetic tree (and for all trees—the so called 'model invariants'), on any number $n$ of leaves. We also provide a precise description of the space of mixtures and linear invariants for the special case of $n = 4$ leaves. By combining techniques from discrete random processes and (multi-) linear algebra, our results build on a classic result that was first established by James Lake (Mol Biol Evol 4:167–191, 1987).

**Keywords** Phylogenetic tree · Markov processes · Linear invariants

✉ Mike Steel
   mike.steel@canterbury.ac.nz

   Marta Casanellas
   marta.casanellas@upc.edu

[1]   Department of Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

[2]   Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

🌀 Springer

**Mathematics Subject Classification** 05C05 · 60J28 · 92D15

## 1 Introduction

Tree-based Markov processes on a discrete state space play a central role in molecular systematics. These processes allow biologists to model the evolution of characters and thereby to develop techniques for inferring a phylogenetic tree for a group of species from a sequence of characters (such as the sites at aligned DNA or amino acid sequences; Felsenstein 2004). Under the assumption that each character evolves independently on the same underlying tree, according to a fixed Markov process, the tree topology can be inferred in a statistically consistent way (i.e. with an accuracy approaching 1 as the number of characters grows) by methods such as maximum likelihood estimation (MLE) (Chang 1996) and techniques based on phylogenetic invariants (Fernández-Sánchez and Casanellas 2016). This holds even though one may not know the values of the other (continuous) parameters associated with the model, which typically relate to the length of the edges, and relative rates of different substitution types.

The assumption that all characters evolve under the same Markov process is a very strong one, and biologists generally allow the underlying process to vary in some way between the characters. For example, a common strategy is to allow characters to evolve at different rates (i.e. the edge lengths are all scaled up or down in equal proportion at each site by a factor sampled randomly from some simple parameterized distribution). In that case, provided the rate distribution is sufficiently constrained, the tree topology can still be inferred in a statistically consistent manner (Allman et al. 2012; Matsen et al. 2008), and by using MLE, or related methods.

However, when this distribution is not tightly constrained, or when edge lengths are free to vary in a more general fashion from character to character then different trees can lead to identical probability distributions on characters (Allman et al. 2012; Steel et al. 1994). In that case, it can be impossible to decide which of two (or more) trees generated the given data, even when the number of characters tends to infinity. In statistical terminology, *identifiability* of the tree topology parameter is lost. For certain types of Markov models, however, identifiability of the tree topology is possible, even in these general settings. These are models for which (i) linear relationships (called 'linear phylogenetic invariants') exist between the probabilities of different characters, and which hold for all values of the other continuous parameters associated with the model (such as edge lengths) and (ii) these invariants can be used to determine the tree topology (at least for $n = 4$ leaves) (Steel 2011; Štefankovič and Vigoda 2007). The first such invariants, which we call *linear topology invariants*, were discovered by James Lake in a landmark paper in 1987 (Lake 1987) for the Kimura 2ST model, and the Jukes–Cantor submodel.

Linear topology invariants were known to exist for Kimura 2ST and Jukes–Cantor models, and the dimension of the corresponding (quotient) linear space had been computed for the Jukes–Cantor model in Fu (1995) and Steel and Fu (1995). It is also known that more general models such as Kimura 3ST or the general Markov model do not admit linear topology invariants (see for example Sturmfels and Sullivant 2005; Casanellas and Fernández-Sánchez 2011). Nevertheless, linear topology invariants

had not been studied for evolutionary models with more than 4 states or for models slightly more general than Jukes–Cantor.

In this paper we extend Lake-type invariants to a more general setting and for another type of process, the 'equal input' model (defined shortly, but it can be regarded as the simplest Markov process that allows different states to have different stationary probabilities). By building also on the approach of Matsen et al. (2008) (which dealt just with the 2-state setting) we investigate the vector space of linear invariants, and describe the space of phylogenetic mixtures on a tree (or trees) under the equal input model once the stationary distribution is fixed. Note that the space of phylogenetic mixtures is dual to the space of phylogenetic invariants, and hence studying one of these spaces translates into results for the other space. This leads to our main results (Theorems 1 and 2) which characterize the space of phylogenetic mixtures across all trees, and on a fixed tree (respectively), along with an algorithm for constructing a basis for the topology invariants. It is worth pointing out that while linear *topology* invariants are relevant for distinguishing distributions arising from mixtures of distributions on particular tree topologies, linear phlylogenetic invariants satisfied by distributions arising from mixtures of distributions on trees evolving under a particular model (*model invariants*) can be used in model selection as in Kedzierska et al. (2012). In brief summary, our main results describe the vector space (and its dimension) of the space of phylogenetic mixtures of the equal input models for any numbers $n$ of leaves and $\kappa$ of states:

– across all trees (Theorem 1) by providing a spanning set of independent points;
– for a fixed tree (Theorem 2); and
– for an infinite state version of the equal input model, known as Kimura's infinite allele model (Proposition 5).

Using the duality between phylogenetic mixtures and linear invariants, in Corollary 1 we compute the dimension of the quotient space of linear topology invariants and describe an algorithm for computing a basis of this space. Note that the dimension of the space of mixtures had already been computed in Casanellas et al. (2012) and in Fu (1995) for the Jukes–Cantor model. Theorem 3 and Corollary 3 provide a more detailed description for trees with $n = 4$ leaves. The case $n = 4$ is of particular interest, since the existence of a set of linear phylogenetic invariants for this case and which, collectively, suffice to identify the tree topology means that there also exist informative linear phylogenetic invariants that can identify any fully-resolved (binary) tree topology on any number of leaves. This follows from the well-known fact that any binary tree topology is fully determined by its induced quartet trees (for details and references, see Semple and Steel 2003).

We also establish various other results along the way, including a 'separability condition' from which a more general description of Lake-type invariants follows (Proposition 3). We begin with some standard definitions, first for Markov processes on trees, and then for the equal input model, which we show is formally equivalent to a random cluster process on a tree (Proposition 2). We then develop a series of preliminary results and lemmas that will lead to the main results described above.

## 2 Markov processes on trees

Given a tree $T = (V, E)$ with leaf set $X$, a *Markov process on $T$* with state space $S$ is a collection of random variables $(Y_v : v \in V)$ taking values in $S$, and which satisfies the following property. For each interior vertex $v$ in $T$, if $V_1, \ldots, V_m$ are the sets of vertices in the connected components of $T - v$ then the $m$ random variables $W_i = (Y_v : v \in V_i)$ are conditionally independent given $Y_v$.

Equivalently, if we were to direct all the edges away from some (root) vertex, $v_0$, then this condition says that conditional on $Y_v$ (for an interior vertex $v$ of $T$) the states in the subtrees descended from $v$ are independent of each other, and are also independent of the states in the rest of the tree.

A Markov process on $T$ is determined entirely by the probability distribution $\pi$ at a root vertex $v_0$, and the assignment $e \mapsto P^{(e)}$, that associates a transition matrix with each edge $e = (u, v)$ of $T$ (the edge is directed away from $v_0$). Matrix $P^{(e)}$ has row $\alpha$ and column $\beta$ entry equal to $P^{(e)}_{\alpha\beta} := \mathbb{P}(Y(v) = \beta | Y(u) = \alpha)$, and so each row sums to 1. If stochastic vector $\pi$ has the property that $\pi = \pi P^{(e)}$ for every edge $e$ of $T$, then $\pi$ is said to be a *stationary distribution* for the process. A *phylogenetic model* is a Markov process on a tree where the transition matrices are required to belong to a particular class $\mathcal{M}$.

In this paper we will be concerned with trees in which the set $X$ of leaves are labelled, and all non-leaf (interior) vertices are unlabelled and have degree at least three; these are called *phylogenetic $X$-trees* (Semple and Steel 2003). A tree with a single interior vertex is called a *star*, while a tree for which every interior vertex has degree three is said to be *binary*. We will write $ab|cd$ for the binary tree on four leaves (a *quartet tree*) that has an edge separating leaves $a$, $b$ from $c$, $d$. A function $\chi : X \to S$ is called a *character* and any Markov process on a tree with state space $S$ induces a (marginal) probability distribution on these characters. An important algebraic feature of this distribution is that the probability of a character $\mathbb{P}(\chi)$ under a Markov process on $T$ is a polynomial function of the entries in the transition matrices.

### 2.1 The equal input model

The *equal input model* ($EI$) for a set $S$ of $\kappa$ states is a particular type of Markov process on a tree, defined as follows. Given a root vertex $v_0$ let $\pi$ be a distribution of states at $v_0$ and for each (directed) edge $e = (u, v)$ (directed away from $v_0$). In the $EI$ model, each transition matrix $P^{(e)}$ has the property that for some value $\theta_e \in [0, 1]$ and all states $\alpha, \beta \in S$ with $\alpha \neq \beta$ we have:

$$P^{(e)}_{\alpha\beta} = \pi_\beta \cdot \theta_e. \tag{1}$$

We shall assume that the distribution $\pi$ is strictly positive throughout the paper.

This model generalizes the familiar *fully symmetric model* of $\kappa$ states (such as the 'Jukes–Cantor model', when $\kappa = 4$) to allow each state to have its own stationary probability. In the case $\kappa = 4$ with $S$ equal to the four nucleotide bases, the model is known as the *Felsenstein 1981 model*. The defining property of the model is that the

probability of a transition from $\alpha$ to $\beta$ (two distinct states) is the same, regardless of the initial state $\alpha (\neq \beta)$.

**Lemma 1** *The following properties hold for the equal input model.*

(i) $P_{\alpha\alpha}^{(e)} = 1 - \theta_e + \pi_\alpha \theta_e$.

(ii) *$\pi$ is a stationary distribution for each vertex $v$ of the $T$ (i.e. $\mathbb{P}(Y(v) = \alpha) = \pi_\alpha$).*

(iii) *The process is time-reversible (i.e. for each edge $e$, $\pi_\alpha P_{\alpha\beta}^{(e)} = \pi_\beta P_{\beta\alpha}^{(e)}$).*

(iv) *If $p$ is the probability that the ends of $e$ receive different states under the EI model, then $p = (1 - \sum_\alpha \pi_\alpha^2)\theta_e$.*

(v) *The process is multiplicatively closed. In other words, $(P^{(e)} P^{(e')})_{\alpha\beta} = \pi_\beta \theta$, where $\theta = 1 - (1 - \theta_e)(1 - \theta_{e'})$.*

*Proof* For (i), $P_{\alpha\alpha}^{(e)} = 1 - \sum_{\beta \neq \alpha} P_{\alpha\beta}^{(e)} = 1 - \theta_e \sum_{\beta \neq \alpha} \pi_\beta = 1 - \theta_e(1 - \pi_\alpha)$. For (ii), it suffices to show that if $(u, v)$ is a directed edge and $u$ has stationary distribution $\pi$ then $v$ does too. But

$$\mathbb{P}(Y(v) = \beta) = \sum_\gamma \pi_\gamma P_{\gamma\beta}^{(e)} = \pi_\beta P_{\beta\beta}^{(e)} + \sum_{\gamma \neq \beta} \pi_\gamma P_{\gamma\beta}^{(e)} = \pi_\beta.$$

For (iii), the result clearly holds if $\alpha = \beta$ so suppose $\alpha \neq \beta$. Then

$$\pi_\alpha P_{\alpha\beta}^{(e)} = \pi_\alpha (\pi_\beta \theta_e) = \pi_\beta (\pi_\alpha \theta_e) = \pi_\beta P_{\beta\alpha}^{(e)}.$$

For (iv),

$$p = \sum_\alpha \pi_\alpha \sum_{\beta \neq \alpha} P_{\alpha\beta}^{(e)} = \sum_\alpha \pi_\alpha \sum_{\beta \neq \alpha} \pi_\beta \theta_e,$$

which simplifies for the expression in (iv). Property (v) is left as an exercise.

For an equal input model, the transition matrix $P^{(e)}$ has eigenvalue $1 - \theta_e$ with multiplicity $k - 1$ (and eigenvalue 1 with multiplicity 1). Also, for fixed $\pi$ the matrices $P^{(e)}$ commute, as they can be simultaneously diagonalized by a fixed matrix (which depends on $\pi$). Equal input models with also have a continuous realisation with rate matrix $Q$ defined by its off-diagonal entries as follows:

$$Q_{\alpha\beta} = \pi_\beta, \quad \text{for all } \alpha, \beta \in S, \alpha \neq \beta$$

(the diagonal entries are determined by the requirement that each row of $Q$ sums to 0). Then $P^{(e)} = \exp(Qt)$ for $t = -\ln(1 - \theta_e)$, and so $\theta_e = 1 - e^{-t}$. In the case where $\pi$ is uniform, the EI model reduces to the fully symmetric model in which all substitution events have equal probability.

One feature of the EI model, that fails for most other Markov processes on trees, is the following. Let $\sigma$ be any partition of the state space $S$, and for a state $s \in S$ let $[s]$ denote the corresponding block of $\sigma$ containing $s$. Then for an EI process $Y$ on the set $V$ of vertices of a phylogenetic tree $T$, let $\tilde{Y}$ be the induced stochastic process on $V$, defined by $\tilde{Y}(v) = [Y(v)]$ for all vertices $v$ of $T$.

**Proposition 1** *For any EI model with parameters $\pi$ and $\{\theta_e\}$, and any partition $\sigma$ of $S$, $\tilde{Y}$ is also an EI Markov process on $T$, with parameters $\tilde{\pi}$ and $\{\theta_e\}$, where for each block $B$ of $\sigma$, $\tilde{\pi}_B := \sum_{\beta \in B} \pi_\beta$.*

*Proof* By Theorem 6.3.2 of Kemeny and Snell (1976), the condition for $\tilde{Y}$ to be a Markov process is that it satisfies a 'lumpability' criterion that for any two choices $\alpha, \alpha' \in A \in \sigma$, and block $B \in \sigma$,

$$\mathbb{P}(Y(v) \in B | Y(u) = \alpha) = \mathbb{P}(Y(v) \in B | Y(u) = \alpha').$$

For each $B \neq A$, this last equality is clear from (1), and since $\mathbb{P}(Y(v) \in A | Y(v) = \alpha) = 1 - \sum_{B \in \sigma, B \neq A} \mathbb{P}(Y(v) \in B | Y(u) = \alpha)$ the criterion also holds for the case $B = A$. Finally, for $B \neq A$, $\mathbb{P}(\tilde{Y}(v) = B | \tilde{Y}(u) = A) = \sum_{\beta \in B} (\pi_\beta \theta_e) = \tilde{\pi}_B \theta_e$. $\qquad \square$

### 2.2 A useful lemma

For results to come the following lemma, and its corollary will be helpful.

**Lemma 2** *For variables $x_1, x_2, \ldots, x_r$, consider polynomials $f_0(\mathbf{x}), \ldots, f_M(\mathbf{x}) \in \mathbb{R}[x_1, \ldots, x_r]$ of the form*

$$f_i(\mathbf{x}) = \sum_{A \subseteq [r]} c_A^{(i)} \prod_{j \in A} x_j, \quad c_A^{(i)} \in \mathbb{R}.$$

(i) *Then $f_0 \equiv 0$ (i.e. $c_A^{(0)} = 0$ for all $A \subseteq [r]$) if and only if for any $t \neq 0$, $f_0(\mathbf{x}) = 0$ for all $\mathbf{x} \in \{0, t\}^r$.*

(ii) *Let $f = (f_1, \ldots, f_M) : \mathbb{R}^r \to \mathbb{R}^M$ and let $L : \mathbb{R}^M \to \mathbb{R}$ be a linear map. Define an equivalence relation among the elements of $\{0, 1\}^r$ by $\mathbf{x} \sim \mathbf{x}'$ if $f(\mathbf{x}) = f(\mathbf{x}')$, and let $\mathbf{x}_1, \ldots, \mathbf{x}_s$ be representatives of these equivalence classes. We call $q_i = f(\mathbf{x}_i)$, $i = 1, \ldots, s$. Then $L(f(\mathbf{x})) = 0$ for all $\mathbf{x} \in \mathbb{R}^r$ if and only if $L(q_j) = 0$ for $j = 1, \ldots, s$.*

*Proof* (i) The 'only if' part holds automatically; for the 'if' direction, given any subset $B$ of $[r]$, let $h(B) = h(\mathbf{x}^B)$ where $x_i^B = t$ if $i \in B$ and $x_i^B = 0$ otherwise. Then $h(B) = 0$ by hypothesis, and $h(B) = \sum_{A \subseteq B} c_A t^{|A|}$, by definition. Applying the (generalized) principle of inclusion and exclusion it follows that, for each $A \subseteq [n]$, $c_A t^{|A|} = \sum_{B \subseteq A} (-1)^{|A-B|} h(B) = 0$, so $c_A = 0$.

(ii) The map $h = L \circ f$ satisfies the hypotheses of (i), hence $L(f(\mathbf{x})) = 0$ for all $\mathbf{x}$ if and only if $L(f(\mathbf{x})) = 0$ for all $\mathbf{x} \in \{0, 1\}^r$. Then the statement follows immediately due to the definition of the equivalence relation. $\qquad \square$

In what follows we will use this lemma to check linear relations among the character probabilities.

In the $EI$ model, once we fix $\pi$, the probability $\mathbb{P}_T(\chi | \Theta)$ of observing a character at the leaves of $T$ satisfies the hypotheses of the corollary with $r$ equal to the number
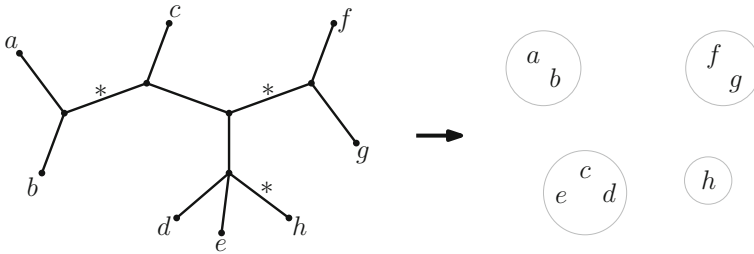
**Fig. 1** Cutting the three edges marked * in the tree on the *left* leads to the partition of $X$ shown at *right*. Under the random cluster model these four blocks are independently assigned states from the distribution $\pi$

of edges and variables in $\Theta = \{\theta_e\}_{e \in E(T)}$. Indeed, by Lemma 1 (i), any entry of the transition matrix $P^{(e)}$ is a linear function of $\theta_e$ and hence the expression

$$\mathbb{P}_T(\chi | \Theta) = \sum_{(s_v)_v \in S^{\text{Int}(T)}} \pi_{s_{v_0}} \prod_{(u,w) \in E(T)} P^{(e)}_{s_u, s_w} \tag{2}$$

(where the sum is over the states at the set $\text{Int}(T)$ of interior vertices of $T$ and subject to the convention that $s_w = \chi(l)$ if $w$ is the leaf $l$) satisfies the hypotheses of Lemma 2.

*Remark 1* Lemma 2 can be slightly modified to accommodate substitution matrices with more parameters as it was done in Fu (1995).

### 2.3 The equal input model as a random cluster model

Our alternative description of the $EI$ model is as an instance of the (finite) *random cluster model* (briefly *RC*) on trees (this phrase is also used to study processes on graphs, such as the 'Ising model' in physics). For an unrooted phylogenetic tree with leaf set $[n]$, each edge $e$ of $T$ is cut independently with probability $p_e$. The leaves in each connected component of the resulting disconnected graph are then all assigned the same state $s$ with probability $\pi_s$, independently of assignments to the other components (see Fig. 1). More precisely, for any binary function $g : E(T) \to \{0, 1\}$, define $C(g)$ to be the set of connected components in $T \setminus \{e \in E(T) | g(e) = 1\}$. Then the probability $\mathbb{P}_T(\chi | \{p_e\}_e)$ of observing a character $\chi$ at the leaves of $T$ under the *RC* model is

$$\sum_{g : E(T) \to \{0,1\}} \mathbb{P}(\chi | g) p_e^{g(e)} (1 - p_e)^{1 - g(e)} \tag{3}$$

where $\mathbb{P}(\chi | g)$ is 0 if $\chi(i) \neq \chi(j)$ for some leaves $i$, $j$ in the same connected component in $C(g)$ and is equal to $\prod_{c \in C(g)} \pi_{\chi_c}$ otherwise (where $\chi_c$ denotes the value of $\chi$ at the leaves of $T$ that are in $c$). In particular, the *RC* model also satisfies the hypotheses of Lemma 2.

**Proposition 2** *The EI model with parameters $\pi$ and $\{\theta_e\}$ produces an identical probability distribution on characters as the random cluster model in which $p_e = \theta_e$ for each edge $e$ of $T$.*

*Proof* For the two models the probability of a given character (given by Eqs. (2) and (3)) satisfies the conditions required by Lemma 2 (ii), and so we can use it with $M = 2$ and $L$ the difference between the probability of a given character by the two models. Therefore, it suffices to show that the two models produce the same probability distribution on characters whenever $\theta_e = 1$ for all $e \in F$ and $\theta_e = 0$ of all edges $e$ of $T$ not in $F$ (for all possible choices of subset $F \in E(T)$). Given $F$, notice that if $\theta_e = 1$ for a directed edge $e = (u, v)$ of $T$ in the $EI$ model, then $P^{(e)}_{\alpha\beta} = \pi_\beta$ for <u>all</u> $\beta \in S$, including $\beta = \alpha$. In other words, when $\theta_e = 1$ for $e = (u, v)$, the state at $v$ is completely independent of the state at $u$. This is equivalent to cutting the edge and assigning a random state according to the distribution $\pi$ to $v$, and thereby to all the other vertices of $T$ for which there is a path to $v$ that does not cross another edge in $F$ (since $P^{(e)}$ is the identity matrix on those edges); this is just the process described by the random cluster model.

## 3 Linear phylogenetic invariants in phylogenetics

**Definition 1** Consider a phylogenetic model $\mathcal{M}$ with state space $S$ on a phylogenetic tree $T$ with $n$ leaves. A *phylogenetic invariant* of a tree $T$ under the model $\mathcal{M}$ is a polynomial $f$ in $S^n$ indeterminates that vanishes on any distribution $\mathbb{P}_{T,\Theta}$ that arises under the phylogenetic model $\mathcal{M}$ (that is, $f(p) = 0$ if $p = \mathbb{P}_{T,\Theta}$, for any set $\Theta$ of transition matrices and distribution at the root vertex).

We say that a polynomial in $S^n$ coordinates is a *model invariant* if it is a phylogenetic invariant for any tree on $n$ leaves under the phylogenetic model $\mathcal{M}$. A phylogenetic invariant of a tree $T$ that is not a model invariant is called a *topology invariant*.

A phylogenetic invariant is a *linear phylogenetic invariant* (resp. *linear model invariant*, *linear topology invariant*) if each monomial involves exactly one indeterminate and has degree 1. Note that this implies that the polynomial is homogeneous (the independent term is 0). There are phylogenetic invariants of degree 1 that are not homogeneous, for example the *trivial* phylogenetic invariant that arises from the observation that in a distribution all coordinates must sum to one. However, taking this trivial invariant into account, any other phylogenetic invariant of degree 1 can be rewritten as a homogeneous phylogenetic invariant of degree 1 (indeed, $\sum_i a_i x_i + a$ is a phylogenetic invariant if and only if $\sum_i (a_i + a) x_i$ is a phylogenetic invariant). This is why we only call *linear* phylogenetic invariants those that are homogeneous of degree 1. The sets of linear model invariants and linear phylogenetic invariants of a tree $T$ are vector spaces.

Linear phylogenetic invariants are of particular interest since they hold even if the process changes from character to character (provide it stays within the model for which the invariant is valid). An important early example of linear phylogenetic invariants were discovered by Lake (1987). In this paper, we first provide a new and more general version of Lake's invariants. It is the first time that linear topology invariants

are given for non-uniform stationary distributions and for models on any number of states, provided that they satisfy what we call the Partial Separability condition (see below). It is worth noting that in our Lake-type invariants the stationary distribution is assumed to be known.

For any phylogenetic $X$-tree, $T$ (not necessarily binary), and an interior vertex $v$ of $T$ consider the disconnected graph $T - v$. Let $t$ and $t'$ be two of the trees incident with $v$.

Suppose that a Markov process $Y$ on $T$ takes values in state space $S$. For any state $s$ of $S$ write $Y(t) = s$ if all the leaves of $T$ that are in $t$ are in state $s$ (similarly for $t'$). Consider the following property.

(PS) *Partial separability.* For some interior vertex $v$, and for some subset $\{a_1, a_2, b_1, b_2\}$ of four distinct elements of $S$ one has

$$\mathbb{P}(Y(t) = a_i | Y(v) = s) = \pi(a_i)c, \quad \text{when } s \in S - \{a_1, a_2\}, \ i = 1, 2;$$

and

$$\mathbb{P}(Y(t') = b_j | Y(v) = s) = \pi(b_j)d, \quad \text{when } s \in S - \{b_1, b_2\}, \ j = 1, 2.$$

Here $c$ and $d$ are arbitrary functions dependent on the tree and associated parameters (but not the states) and $\pi$ is an arbitrary function of the states such that $\pi(a_i) \neq 0$, $\pi(b_i) \neq 0$, $i = 1, 2$ (for various models with $\pi$ given by the stationary distribution).

Partial separability is satisfied by various models. For example, when $|S| = 4$, it holds for the Kimura 2-ST model (and hence the Jukes–Cantor model) by taking $\{a_1, a_2\} = \{A, G\}$ (purines) and $\{b_1, b_2\} = \{C, T\}$ (pyrimidines), in which case $\pi(a_i) = \pi(b_i) = \frac{1}{4}$ for $i = 1, 2$. The property also holds for the fully symmetric model on any number of states. Moreover, the property holds for the $EI/RC$ model on any number of states if $t$ and $t'$ are single leaves. The partial separability condition should be viewed as an algebraic constraint rather than as a natural condition that one might expect to hold for most evolutionary models. For instance it, is not a natural property satisfied by evolutionary models and, for instance, it is not satisfied for the $EI/RC$ model if $t$ or $t'$ are not single leaves.

Let $\mathcal{E}$ be *any* event that involves the states at the leaves of $T$ not in $t$ or $t'$. For example, if $y$ and $y'$ are leaves of $T$ not in $t$ or $t'$ then $\mathcal{E}$ might be the event that $Y(y) = s$ and $Y(y') = s'$ for some particular $s, s' \in S$.

Let us write $p_{\mathcal{E}ij}$ for the probability of the three-way conjunction $\mathcal{E} \wedge \{Y(t) = a_i\} \wedge \{Y(t') = b_j\}$. Notice that $p_{\mathcal{E}ij}$ is a sum of probabilities of various characters (i.e. a marginal distribution). Let

$$\tilde{p}_{\mathcal{E}ij} = \frac{1}{\pi(a_i)\pi(b_j)} \cdot p_{\mathcal{E}ij} \quad \text{and let } \Delta := \tilde{p}_{\mathcal{E}11} + \tilde{p}_{\mathcal{E}22} - \tilde{p}_{\mathcal{E}12} - \tilde{p}_{\mathcal{E}21}.$$

**Proposition 3** (Lake-type invariants) *If a Markov process on $T$ satisfies the partial separability condition (PS), then $\Delta = 0$.*

*Proof* By the Markov property,

$$p_{\mathcal{E}ij} = \sum_s \mathbb{P}(Y(v) = s) \cdot \mathbb{P}(\mathcal{E}|Y(v) = s) \cdot \mathbb{P}(Y(t) = a_i|Y(v) = s)$$
$$\cdot \mathbb{P}(Y(t') = b_j|Y(v) = s).$$

Let $r_{ij} = \pi(a_i) \cdot \pi(b_j)$, and let

$$\Delta_s = r_{22}p_1p_1' + r_{11}p_2p_2' - r_{12}p_2p_1' - r_{21}p_1p_2',$$

where $p_i = \mathbb{P}(Y(t) = a_i|Y(v) = s)$, and $p_j' = \mathbb{P}(Y(t') = b_j|Y(v) = s)$. Then we can write

$$\Delta = \frac{1}{\pi(a_1)\pi(a_2)\pi(b_1)\pi(b_2)} \sum_s \mathbb{P}(Y(v) = s) \cdot \mathbb{P}(\mathcal{E}|Y(v) = s) \cdot \Delta_s.$$

Thus it suffices to show that $\Delta_s = 0$ for all $s$.

We consider three cases: (i): $s \in \{a_1, a_2\}$, (ii) $s \in \{b_1, b_2\}$ and (iii) $s \in S - \{a_1, a_2, b_1, b_2\}$.

In Case (i), suppose $s = a_i$. Then $p_1' = \pi(b_1)d$ and $p_2' = \pi(b_2)d$, and so

$$\Delta_s = d[p_1r_{22}\pi(b_1) + p_2r_{11}\pi(b_2) - p_2r_{12}\pi(b_1) - p_1r_{21}\pi(b_2))].$$
$$= dp_1[r_{22}\pi(b_1) - r_{21}\pi(b_2))] + dp_2[r_{11}\pi(b_2) - r_{12}\pi(b_1)] = 0 + 0 = 0.$$

Case (ii) is similar. In Case (iii), $p_i p_j' = r_{ij}cd$ and so

$$\Delta_s = cd[r_{22}r_{11} + r_{11}r_{22} - r_{12}r_{21} - r_{21}r_{12}] = 0.$$

*Example 1* When we take $t$ and $t'$ single leaves, the $EI/RC$ model satisfies the (PS) property and Lemma 3 can be applied. If the stationary distribution $\pi$ is fixed, then $\Delta$ gives rise to two types of linear phylogenetic invariants for the quartet tree 12|34,

$$H_1 : \frac{\mathsf{x}_{xyxy}}{\pi(x)\pi(y)} + \frac{\mathsf{x}_{xyzw}}{\pi(z)\pi(w)} - \frac{\mathsf{x}_{xyzy}}{\pi(z)\pi(y)} - \frac{\mathsf{x}_{xyxw}}{\pi(x)\pi(w)}$$

$$H_2 : \frac{\mathsf{x}_{xyyx}}{\pi(x)\pi(y)} + \frac{\mathsf{x}_{xywz}}{\pi(z)\pi(w)} - \frac{\mathsf{x}_{xyyz}}{\pi(z)\pi(y)} - \frac{\mathsf{x}_{xywx}}{\pi(x)\pi(y)}$$

(here $\mathsf{x}_{\chi_1\chi_2\chi_3\chi_4}$ is the coordinate that corresponds to $\mathbb{P}_T(\chi_1\chi_2\chi_3\chi_4)$). To see how these follow from Proposition 3, for $H_1$ take $x = a_1$, $y = b_1$, $z = a_2$, $w = b_2$ and let $\mathcal{E}$ be the event that $Y(1) = a_1$ and $Y(2) = b_1$; for $H_2$ take $x = b_1$, $y = a_1$, $z = b_2$, $w = a_2$ and let $\mathcal{E}$ be the event that $Y(1) = b_1$ and $Y(2) = a_1$. Note that these are topology invariants because the first is not a phylogenetic invariant for the quartet 13|24 while the second is not a phylogenetic invariant for 14|23.

## 4 Generating linear invariants for the $RC/EI$ model on $\kappa$ states

### 4.1 Combinatorial concepts and terminology

Let $T$ be a phylogenetic $X$-tree, $X = [n]$, and consider a Random Cluster model (or Equal Input model) on $T$, with stationary distribution $\pi$ on a set $S$ of $\kappa$ states. Henceforth we assume that $\pi$ is fixed and it is positive, that is, $\pi_s \neq 0 \; \forall s \in S$. We denote by $e_i$ the pendant edge incident with leaf $i$. A character $\chi : [n] \to S$ shall be denoted as $\chi = \chi_1 \ldots \chi_n$ if $\chi_i = \chi(i)$ for $i = 1, \ldots, n$. We let $Ch(n, \kappa)$ to be the set of characters on $[n]$ for a fixed state space $(S)$ of size $\kappa$ and denote by $N$ its cardinality ($N = \kappa^n$). We think of a distribution $\mathbb{P}_{T,\Theta}$ on the set of characters under the $RC$ model on $T$ as a vector of $Ch(n, \kappa)$ coordinates and therefore lying in the real vector space with coordinates $x_\chi$, $\chi \in Ch(n, \kappa)$ (the point $\mathbb{P}_{T,\theta}$ has coordinate $x_\chi$ equal to $\mathbb{P}_T(\chi | \Theta)$).

Let $F$ be a *subforest* of $T$, that is, a subgraph comprised of a collection of vertex disjoint trees $\{T_1, \ldots, T_r\}$ such that the only nodes of degree $\leq 1$ in $T_i$ are leaves of $T$ (we allow $T_i$ to be formed by only one leaf and we allow $F = \{T\}$ also). We say that a subforest $F = \{T_1, \ldots, T_r\}$ is a *full* subforest if $\cup_i \mathcal{L}(T_i) = X$; we let $\mathcal{F}_T$ be the set of full subforests of $T$. For a full subforest $F$, we define $\Theta_F$ to be the following collection of edge parameters under the $RC$ model: $\theta_e = 0$ if $e \in E(T_i)$ for some $T_i \in F$ and $\theta_e = 1$ for all other edges $e$. We denote by $\sigma(F)$ the partition that $F$ describes on $[n]$, that is, two leaves are in the same block of $\sigma(F)$ if they lie in the same subtree of $F$. The full subforest formed by singletons will be called the *trivial* subforest.

Given a character $\chi$, we define $\sigma(\chi)$ to be the partition $\{S_1, \ldots, S_l\}$ of the set of leaves defined according to "two leaves $i, j$ are in the same block of the partition if $\chi_i = \chi_j$". Note that given a full subforest $F = \{T_1, \ldots, T_r\}$ of $T$ and a character $\chi$, $\mathbb{P}_T(\chi | \Theta_F)$ is zero if $\sigma(F)$ does not refine $\sigma(\chi)$ and is equal to $\prod_{i=1}^{r} \pi_{s_i}$ otherwise (here $s_i$ stands for the value of $\chi$ at the leaves of $T_i$).

For any partition $\sigma$ of $[n]$, and any phylogenetic tree $T$ on $[n]$, we say that $\sigma$ is *convex* on $T$ (or *compatible* with $T$) if the collection of induced subtrees $\{T[B] : B \in \sigma\}$ are vertex disjoint (here $T[B]$ is the minimal connected subgraph (subtree) of $T$ containing the leaves in $B$). Let $co(T)$ be the set of partitions of $[n]$ that are convex on $T$. There is a natural correspondence between full subforests of $T$ and convex partitions on $T$ that associates to each partition $\sigma \in co(T)$ the full subforest $F_T(\sigma) = \{T[B] : B \in \sigma\}$. Therefore, the number of full subforests of a tree $T$ is equal to $|co(T)|$, $|\mathcal{F}_T| = |co(T)|$. When $T$ is a binary tree, $|co(T)| = F_{2n-1}$ where $F_k$ is the $k$-th Fibonacci number, starting with $F_1 = F_2 = 1$ (see Steel and Fu 1995). By contrast, for a star tree on $[n]$ we have $|co(T)| = 2^n - n$. A partition $\sigma = \{B_1, \ldots, B_k\}$ of $[n]$ is *incompatible* with $T$ if it is not convex on $T$, that is, there exist two blocks $B_i$ and $B_j$ from $\sigma$ for which $T[B_i]$ and $T[B_j]$ share at least one vertex. A *singleton block $B$* of $\sigma$ is a block of size 1. The number of partitions of $[n]$ is known as the *Bell number $B_n$*.

Finally, let $Inc(T)$ be the set of partitions of $[n]$ that are not convex on $T$ (i.e. they are 'incompatible' with $T$). Thus $|Inc(T)| = B_n - |co(T)|$.

### 4.2 Results

**Lemma 3** (a) *Let $\Theta$ be a collection of parameters $(\theta_e)_{e \in E(T)}$ such that $\theta_e$ is either 0 or 1 for all $e \in E(T)$. Then there exists a unique full subforest $F \in \mathcal{F}_T$ such that $\mathbb{P}_{T,\Theta} = \mathbb{P}_{T,\Theta_F}$.*

(b) *A degree 1 polynomial $\sum_\chi \lambda_\chi x_\chi$ is a linear phylogenetic invariant for a tree $T$ if and only if*

$$\sum_\chi \lambda_\chi \mathbb{P}_T(\chi|\Theta_F) = 0$$

*for any full subforest $F \in \mathcal{F}_T$.*

*Proof* (a) We first prove that two full subforests $F$ and $G$ satisfy $\mathbb{P}(\chi|\Theta_G) \neq \mathbb{P}(\chi|\Theta_F)$ for some $\chi$ if $F \neq G$. As $F, G$ are full subforests, they are different if and only if they induce different partitions $\sigma(F), \sigma(G)$ on the set of leaves. Then there exists an edge $e_0$ such that $e_0$ is compatible with $\sigma(F)$ (i.e, $\sigma(F)$ refines the bipartition induced by $e_0$) but is not compatible with $\sigma(G)$ (or the other way around). If $\chi$ is the character that assigns state $x$ at the leaves of one connected component of $T - e_0$ and state $y \neq x$ at the leaves of the other component, then $\mathbb{P}(\chi|\Theta_G) = 0$ while $\mathbb{P}(\chi|\Theta_F)$ is not zero.

Given $\Theta$, let $A$ be the set of edges $e$ in $T$ such that $\theta_e = 1$. Let $\sigma(T \setminus A)$ be the partition induced on $X$ when removing all edges in $A$ (if an edge in $A$ is a pendant edge, then removing it means that we separate the corresponding leaf). If $F$ is the subforest $F_T(\sigma(T \setminus A))$, then we have $\mathbb{P}_{T,\Theta} = \mathbb{P}_{T,\Theta_F}$.

(b) This follows from part (a) and Lemma 2 (ii).

Let $\Theta$ be a collection of edge parameters on a tree $T$ evolving under the $RC$ model. For a site character $\chi$, we define

$$\tilde{p}_\chi^T(\Theta) = \frac{\mathbb{P}_T(\chi|\Theta)}{\pi_{\chi_1}\pi_{\chi_2}\ldots\pi_{\chi_n}}.$$

We call $\tilde{x}_\chi$ the corresponding coordinates: $\tilde{x}_\chi = \frac{x_\chi}{\pi_{\chi_1}\pi_{\chi_2}\ldots\pi_{\chi_n}}$.

**Lemma 4** *We say that two characters $\chi$ and $\chi'$ are* equivalent, *$\chi \equiv \chi'$, if $\sigma(\chi) = \sigma(\chi')$ and $\chi_i = \chi_i'$ for any leaf $i$ that belongs to a block of the partition of cardinality greater than or equal to 2. Let $\chi$, $\chi'$ be two characters on the set $X = [n]$.*

(a) *If $\chi \equiv \chi'$ then $\tilde{x}_\chi - \tilde{x}_{\chi'}$ is a linear model invariant.*

(b) *If $\pi$ is not invariant by any permutation of the set of states, then for any tree $T$ the equality $\tilde{p}_\chi^T(\Theta) = \tilde{p}_{\chi'}^T(\Theta)$ for every $\Theta$ implies that $\chi \equiv \chi'$ (i.e. in this case every linear phylogenetic invariant of type $\tilde{x}_\chi - \tilde{x}_{\chi'}$ satisfies $\chi \equiv \chi'$).*

*Proof* (a) Let $\chi$ and $\chi'$ be two equivalent characters, let $\sigma$ be $\sigma(\chi) = \sigma(\chi')$, and let $T$ be any $X$-tree. According to Lemma 3 (b) we need to check that $\tilde{p}_\chi(\Theta_F) = \tilde{p}_{\chi'}(\Theta_F)$ for any $F = \{T_1, \ldots, T_r\} \in \mathcal{F}_T$.

If $\sigma(F)$ does not refine $\sigma$, then $\mathbb{P}_T(\chi|\Theta_F)$ and $\mathbb{P}_T(\chi'|\Theta_F)$ are zero and we are done.

If $\sigma(F)$ does refine $\sigma$, then $\mathbb{P}_T(\chi|\Theta_F) = \pi_{s_1} \ldots \pi_{s_r}$ where $s_i$ is the value of $\chi$ at the leaves of $T_i$ (note that we may have $s_i = s_j$). Therefore $\tilde{p}_\chi^T(\Theta_F) = \frac{1}{\pi_{s_1}^{n_1-1} \ldots \pi_{s_r}^{n_r-1}}$ where $n_i = |\mathcal{L}(T_i)|$. As $\sigma(F)$ refines $\sigma(\chi) = \sigma(\chi')$ and the states of $\chi$ and $\chi'$ coincide for any block of $\sigma$ of size $\geq 2$, the states of $\chi$ and $\chi'$ also coincide at the leaves of $T_i$ if $n_i \geq 2$. Therefore, $\tilde{p}_\chi^T(\Theta_F) = \tilde{p}_{\chi'}^T(\Theta_F)$.

As for (b), assume that $\pi$ is not invariant by any permutation of the set of states (i.e. $\pi_s = \pi_t$ if and only if $s = t$). Assume that for a tree $T$ we have $\tilde{p}_\chi^T(\Theta_T) = \tilde{p}_{\chi'}^T(\Theta_T)$ for any collection of edge parameters $\Theta_T$. Then, for each block $B_i$ of $\sigma(\chi)$ of size $b_i$ greater or equal than 2 consider the forest $F_i = \{T_{B_i}, \cup_{l \notin B_i}\{l\}\}$, where $T_{B_i}$ is the smallest subtree of $T$ joining the leaves in $B_i$. Then $\tilde{p}_\chi^T(\Theta_{F_i}) = \frac{1}{\pi_{s_i}^{b_i-1}}$ if $s_i$ is the state of $\chi$ at the leaves of $B_i$. By hypothesis this is equal to $\tilde{p}_{\chi'}^T(\Theta_{F_i})$. But $\tilde{p}_{\chi'}^T(\Theta_{F_i})$ is zero if $\sigma(\chi')$ does not contain the block $B_i$. Performing the same argument for any block $B_i$ of size $b_i \geq 2$ we obtain $\sigma(\chi) = \sigma(\chi')$. Now for each such block $B_i$ we have $\tilde{p}_\chi^T(\Theta_{F_i}) = \tilde{p}_{\chi'}^T(\Theta_{F_i})$ and hence $\frac{1}{\pi_{s_i}^{b_i-1}} = \frac{1}{\pi_{s_i'}^{b_i-1}}$ if $s_i'$ is the state of $\chi'$ at the leaves of $B_i$. As $b_i \geq 2$, the assumption on $\pi$ implies $s_i = s_i'$. Thus, $\chi$ and $\chi'$ are equivalent characters.

*Remark 2* If $\pi$ is the uniform distribution (i.e we consider the $\kappa$-state fully symmetric model), then we have $\mathbb{P}_T(\chi|\Theta) = \mathbb{P}_T(\chi'|\Theta)$ if and only if $\sigma(\chi) = \sigma(\chi')$. Indeed, in this case if we consider any permutation $g$ of the set of states $S$, the polynomials $x_\chi - x_{g \cdot \chi}$ are linear phylogenetic invariants for any tree (see Casanellas et al. 2012), where $g \cdot \chi$ stands for the corresponding permutation of states at the leaves. But these polynomials can also be rewritten as $x_\chi - x_{\chi'}$ for $\sigma(\chi) = \sigma(\chi')$.

**Examples:** $n = 3$ and $n = 4$

- For $n = 3$, Lemma 4 gives the following. If $\kappa \geq 3$ and we consider three different states $x$, $y$, $z$ and another set of three different states $x'$, $y'$, $z'$, the linear invariants obtained in Lemma 4 are:

$$\tilde{x}_{xyz} - \tilde{x}_{x'y'z'}, \ \tilde{x}_{xxy} - \tilde{x}_{xxz}, \ \tilde{x}_{xyx} - \tilde{x}_{xzx}, \ \tilde{x}_{yxx} - \tilde{x}_{zxx}.$$

- For $n = 4$, Lemma 4 gives the following. If $\kappa \geq 4$ and we consider four different states $x$, $y$, $z$, $w$ and another set of four different states $x'$, $y'$, $z'$, $w'$, the linear phylogenetic invariants of Lemma 4 are:

$$\tilde{x}_{xyzw} - \tilde{x}_{x'y'z'w'}, \ \tilde{x}_{xxyz} - \tilde{x}_{xxy'z'}, \ \tilde{x}_{xxxy} - \tilde{x}_{xxxy'},$$

and the analogous invariants obtained for the other partitions of [4] involving singletons. $\square$

There are several ways to construct linear invariants from smaller trees and a systematic way to find model invariants for certain models with stationary distribution

has been described in Fu and Li (1991). The most immediate one, used already in the quoted paper, uses the following marginalization lemma. If $i$ is a leaf of $T$, we call $T_i$ the tree obtained by removing leaf $i$ and its incident edge, and suppressing the resulting degree-2 vertex if the interior node had degree 3.

**Lemma 5** *Let $i$ be a leaf of a phylogenetic $[n]$-tree $T$ and let $T_i$ be the corresponding tree. Let $l$ be a linear homogeneous map $l : \mathbb{R}^{\kappa^{n-1}} \to \mathbb{R}$ and let $M_i : \mathbb{R}^{\kappa^n} \to \mathbb{R}^{\kappa^{n-1}}$ be the marginalization map at leaf $i$. If $l(p_i) = 0$ for any distribution $p_i$ from a Markov process on the tree $T_i$, then $(l \circ M_i)(p) = 0$ for any distribution $p$ that comes from a Markov process on the tree $T$.*

*Proof* To prove this lemma one just needs to observe that for any distribution $p$ coming form a Markov process on $T$, $M_i(p)$ is a distribution on $T_i$ that comes from the Markov process that at each edge $e$ has the same transition matrices as $e$ had on the tree $T$.

Another construction, which is new, and particular to the $RC/EI$ model is described in the following lemma. This lemma shall be used in Sect. 6 where we provide specific linear invariants for quartet trees.

**Lemma 6** (Extension lemma) *Let $\Delta = \sum_{\chi} a_{\chi} \mathrm{x}_{\chi}$ be a linear invariant for an $[n]$-tree $T$ evolving under the RC model.*

(a) *Let $T'$ be the tree obtained by subdividing an edge of $T$ and attaching a new pendant edge at the newly introduced node. Let $s$ be a new state not involved in $\Delta$ (that is, $a_{\chi} = 0$ if $\chi$ contains $s$). Then,*

$$\sum_{\chi} a_{\chi} \mathrm{x}_{\chi s} \tag{4}$$

*is a linear invariant for $T'$ (where the new leaf is labelled as leaf $n+1$).*

(b) *Let $T'$ be the tree obtained by subdividing an edge of $T$ and attaching a tree $\tilde{T}$ of $m+1$ leaves to the newly introduced node (so that $T'$ has $n+m$ leaves and the newly introduced leaves are labelled from $n+1$ to $n+m$). Let $\mu$ be a character on $m$ leaves for which $a_{\chi} = 0$ if $\chi$ contains some state in $\mu$ (that is, $\Delta$ does not involve the states of $\mu$ at any leaf). Then $\sum_{\chi} a_{\chi} \mathrm{x}_{\chi \mu}$ is a linear phylogenetic invariant for $T'$ (where $\chi \mu$ stands for states $\chi$ at the first $n$ leaves and states $\mu$ at the other $m$ leaves).*

(c) *Suppose $T$ is the star tree, and let $\mu$ be a character on $m$ leaves for which $a_{\chi} = 0$ if $\chi$ contains some state in $\mu$. Then, for the star tree $T'$ with $n+m$ leaves evolving under the RC model, $\sum_{\chi} a_{\chi} \mathrm{x}_{\chi \mu}$ is a linear phylogenetic invariant.*

*Proof* (a) By Lemma 3, we only need to check that (4) vanishes for the distributions generated with $\Theta = \Theta_F$ where $F$ is a full subforest of $T'$. We denote by $\Theta_{F|T}$ the corresponding probabilities at the edges of $T$ and we denote by $\Delta(\Theta_{F|T})$ the value of $\Delta$ evaluated at $\mathbb{P}_{T, \Theta_{F|T}}$.

If $F$ contains a tree with the new edge $e'$ on it, then, for all $\chi$ involved in $\Delta$, we have $\mathbb{P}_{T'}(\chi s | \Theta_F) = 0$ (because $s$ is a state not involved in $\Delta$) and then (4) trivially vanishes. If $F$ does not contain the edge $e'$, then the new leaf is a singleton in $F$. In

this case we have $\mathbb{P}_{T'}(\chi s | \Theta_F) = \pi_s \mathbb{P}_T(\chi | \Theta_{F|T})$. Therefore (4) evaluated at $\mathbb{P}_{T',\Theta_F}$ is $\Delta(\theta_{F|T})$ multiplied by $\pi_s$, so it vanishes as well.

(b) If $\tilde{T}$ is binary, then the addition of $\tilde{T}$ can be obtained by successively adding cherries to $T$. So, assume that we have added one cherry as in (a), so that we have assigned state $s$ to the new leaf $l_{n+1}$, and now we add a new cherry to the edge leading to $l_{n+1}$. Now the new state $s'$ that we consider for the new leaf now can be allowed to be equal to the state $s$ as long as $s'$ differs from the states that appear in $\Delta$. Indeed, if $s' = s$, there might be forests containing the new cherry, but all of them give probability zero for the states appearing in the polynomial except if the forest is formed by the new cherry and other trees. For such a forest $F$ we have $\mathbb{P}(\chi ss | \Theta_F) = \pi_s \mathbb{P}(\chi | \Theta_{F|T})$ and hence the polynomial evaluated at the parameters of this forest is $\Delta(\Theta_{F|T})$ multiplied by $\pi_s$ which vanishes again.

If $\tilde{T}$ is not a binary tree, then it can be also constructed from a binary tree by contracting edges. As for binary tree the polynomial is a phylogenetic invariant, so it is when we contract edges (note that if a polynomial is a phylogenetic invariant for a tree, then it is also a phylogenetic invariant for the tree $T_0$ obtained by contracting one edge $e_0$ because any collection of edge parameters at $T_0$ gives a collection of edge parameters for $T$ by assigning $\theta_{e_0} = 0$).

(c) This follows from (b) by contracting edges.

## 5 Phylogenetic mixtures

So far, we have found some linear polynomials that turn out to be either model invariants or topology invariants. But we were not able to say whether these invariants actually generate the space of linear phylogenetic invariants for a tree $T$. On the other hand, it would be interesting to know whether a distribution where all these linear invariants vanish is actually a linear combination from distributions on a tree or a mixture of trees. To this end, one defines the space of *mixtures* on a tree (Štefankovič and Vigoda 2007).

**Definition 2** Fix a distribution $\pi$ on the set of states. Given a particular tree $T$, we denote by $\mathbb{P}_{T,\Theta}$ the distribution of a $RC$ model with parameters $\pi$, $\Theta$ on $T$. We define the *space of mixtures on $T$* as

$$\mathcal{D}_T^{\pi} = \left\{ p = \sum_i \lambda_i \mathbb{P}_{T,\Theta_i} \,\Big|\, \sum_i \lambda_i = 1 \right\}.$$

If $\mathcal{T}$ is the set of phylogenetic trees on $[n]$, we define *the space of phylogenetic mixtures on $[n]$* as

$$\mathcal{D}^{\pi} = \left\{ p = \sum_i \lambda_i \mathbb{P}_{T_i,\Theta_i} \,\Big|\, \sum_i \lambda_i = 1, \, T_i \in \mathcal{T} \right\}$$

When $\{p_i\}_{i \in I}$ is a set of points in an affine linear space, we denote by $\langle p_i \,|\, i \in I \rangle_a$ the linear span of these points, that is, the set of points $q = \sum_i \lambda_i p_i$ with $\sum_i \lambda_i = 1$

(we put the subindex $a$ in order to distinguish this affine linear span from the usual linear span of vectors). Note that the spaces of phylogenetic mixtures are affine linear varieties,

$$\mathcal{D}_T^\pi = \left\langle p \,\middle|\, p = \mathbb{P}_{T,\Theta} \right\rangle_a , \quad \mathcal{D}^\pi = \left\langle p \,\middle|\, p = \mathbb{P}_{T,\Theta}, \, T \in \mathcal{T} \right\rangle_a ,$$

and both lie inside the hyperplane

$$H = \left\{ \mathrm{x} = (\mathrm{x}_\chi)_\chi \in \mathbb{R}^N \,\middle|\, \sum_{\chi \in Ch(n,\kappa)} \mathrm{x}_\chi = 1 \right\} .$$

Strictly speaking, for applications in phylogenetics it is only relevant to consider points in $\mathcal{D}^\pi$ (or $\mathcal{D}_T^\pi$) that are actually distributions. In other words, one should be mainly interested in convex combinations of the points $\mathbb{P}_{T,\Theta}$:

$$\left\{ p = \sum_i \lambda_i \mathbb{P}_{T,\Theta_i} \,\middle|\, \lambda_i \geq 0, \sum_i \lambda_i = 1 \right\} \quad \text{and}$$

$$\left\{ p = \sum_i \lambda_i \mathbb{P}_{T_i,\Theta_i} \,\middle|\, \lambda_i \geq 0, \sum_i \lambda_i = 1, \, T_i \in \mathcal{T} \right\} .$$

However, as the dimension of a polyhedron is the dimension of its affine hull, we focus on computing the dimension of $\mathcal{D}^\pi$ and $\mathcal{D}_T^\pi$.

For any distribution $\pi$, we denote by $L^\pi$ the vector space of linear model invariants and by $L_T^\pi$ the space of all linear phylogenetic invariants for a tree $T$. The orthogonal subspace of $L^\pi$ (respectively $L_T^\pi$) shall be denoted by $E^\pi$ (respectively $E_T^\pi$), that is, $E^\pi$ is the set of vectors in $\mathbb{R}^N$ where all the linear model invariants vanish and $E_T^\pi$ the set of vectors where all the linear phylogenetic invariants for $T$ vanish (by identifying dual and orthogonal spaces). In other words, $E_T^\pi$ and $E^\pi$ are spanned by the following vectors of distributions:

$$E_T^\pi = \left\langle \mathbf{p} \,\middle|\, \mathbf{p} = \mathbb{P}_{T,\Theta} \right\rangle , \quad E_T^\pi = \left\langle \mathbf{p} \,\middle|\, \mathbf{p} = \mathbb{P}_{T,\Theta}, \, T \in \mathcal{T} \right\rangle .$$

Note that when we use $p \in \mathbb{R}^N$ as a vector, we use the notation $\mathbf{p}$ to distinguish it from its use as an affine point in $\mathbb{R}^N$. Then the following equalities are clear

$$\mathcal{D}_T^\pi = E_T^\pi \cap H , \quad \mathcal{D}^\pi = E^\pi \cap H.$$

Therefore, studying phylogenetic mixtures (on $[n]$ or on a tree) is equivalent to studying linear phylogenetic invariants (only model invariants or together with topology invariants). Note that due to Lemma 3, it is clear that

$$E_T^\pi = \langle \mathbf{p} = \mathbb{P}_{T,\Theta_F} \,|\, F \in \mathcal{F}_T \rangle , \quad E^\pi = \langle \mathbf{p} = \mathbb{P}_{T,\Theta_F} \,|\, T \in \mathcal{T}, \, F \in \mathcal{F}_T \rangle$$

(see also Matsen et al. 2008, Prop. 10).

In this section we compute the dimension of the spaces of phylogenetic mixtures.

### 5.1 Model invariants and phylogenetic mixtures

We fix $n \geq 4$ throughout this section. We call $\Sigma_\kappa$ the set of partitions of $[n]$ of size at most $\kappa$ (note that if $\kappa \geq n$, this is the whole set of partitions of $[n]$). If $\sigma$ is a partition of $[n]$ compatible with trees $T$ and $T'$, and we consider $F = F_T(\sigma)$ and $F' = F_{T'}(\sigma)$, then one has $\mathbb{P}_{T,\Theta_F} = \mathbb{P}_{T',\Theta_{F'}}$. This point will be briefly denoted as $q_\sigma$ (because it does not depend on the chosen tree compatible with $\sigma$). We give the coordinates of the points $q_\sigma$ for $n = 4$ shortly, see Example 2. Note that $\mathcal{D}^\pi = \langle q_\sigma \mid \sigma \in \Sigma_n \rangle_a$, but this spanning set of points are not affine linearly independent if $\kappa \geq n$:

**Theorem 1** *If $\pi$ is a distribution on $\kappa$ states with positive entries, then $\{q_\sigma \mid \sigma \in \Sigma_\kappa\}$ are affine linearly independent points. Moreover, if $\pi$ is the uniform distribution or a generic distribution, or if $\kappa \geq n$, then $\mathcal{D}^\pi$ coincides with $\langle q_\sigma \mid \sigma \in \Sigma_\kappa \rangle_a$ and has dimension $|\Sigma_\kappa| - 1$ (which equals $B_n - 1$ if $\kappa \geq n$).*

The inclusion $\langle q_\sigma \mid \sigma \in \Sigma_\kappa \rangle_a \subseteq \mathcal{D}^\pi$ clearly holds (and if $\kappa \geq n$, the other inclusion is trivial). The idea for the proof of the other inclusion is to use $\mathcal{D}^\pi = E^\pi \cap H$, bound the dimension of $E^\pi$ from above by a quantity $d$ and prove that the set of points $q_\sigma$ span an affine linear variety of dimension $d - 1$. We first need the following lemma.

**Lemma 7** (a) *For any $\kappa$, the set $\{q_\sigma \mid \sigma \in \Sigma_\kappa\}$ is formed by affine linearly independent points for any distribution $\pi$ (with positive entries).*
(b) *If $\pi_U$ is the uniform distribution, then the set of linear model invariants is spanned by the set of polynomials $x_\chi - x_{\chi'}$ for $\sigma(\chi) = \sigma(\chi')$. In particular, the set of vectors $E^{\pi_U}$ where the model invariants vanish has dimension equal to $|\Sigma_\kappa|$.*

*Proof* (a) We need to prove that if we have a linear combination

$$\sum_{\sigma \in \Sigma_\kappa} \lambda_\sigma q_\sigma = 0 \tag{5}$$

with $\sum_\sigma \lambda_\sigma = 0$, then we need to prove that the coefficients $\lambda_\sigma$ are zero. We proceed by induction on $m = \min\{n, \kappa\}$. Note that as all partitions of $[n]$ are of size at most $n$, $\Sigma_\kappa$ equals the set $\Sigma_m$ of partitions of size at most $m$.

If $m = 1$, then $\Sigma_\kappa$ contains a single element and there is nothing to prove. Assume that $m \geq 2$ and consider a linear combination as in Eq. (5).

Note that the coordinate $\tilde{x}_\chi$ of $q_\sigma$ is zero if $\sigma$ does not refine $\sigma(\chi)$. Let $\tilde{x}_\chi$ be a coordinate such that $\sigma(\chi)$ has the maximum size $m$. Then $\tilde{x}_\chi$ is different from zero only for $q_{\sigma(\chi)}$ (because the other points $q_\sigma$ correspond to partitions that do not refine $\sigma(\chi)$). Thus, $\lambda_{\sigma(\chi)} = 0$ and hence in (5) we have $\lambda_\sigma = 0$ for all $\sigma$ of size $m$. Thus, we are left with a linear combination such as

$$\sum_{\sigma \in \Sigma_{m-1}} \lambda_\sigma q_\sigma = 0, \quad \sum_{\sigma \in \Sigma_{m-1}} \lambda_\sigma = 0.$$

The result follows by the induction hypothesis.

(b) For the uniform distribution, each polynomial $x_\chi - x_{\chi'}$ for $\sigma(\chi) = \sigma(\chi')$ is clearly a model invariant (see Remark 2). Thus the set of vectors $E^{\pi_U}$ where these polynomials vanish has dimension less than or equal to $|\Sigma_\kappa|$. The set of points considered in (a) for $\pi_U$ is contained in $E^{\pi_U} \cap H$, and hence (as $H$ is an equation linearly independent with the previous polynomials), the dimension of $E^{\pi_U}$ is $|\Sigma_\kappa|$. It follows that the inclusion $E^{\pi_U} \subseteq \{x \in \mathbb{R}^N | x_\chi = x_{\chi'} \text{ if } \sigma(\chi) = \sigma(\chi')\}$ is actually an equality and the set of model invariants is spanned by the polynomials $x_\chi - x_{\chi'}$ for $\sigma(\chi) = \sigma(\chi')$.

Now we are ready to prove the theorem.

*Proof of Theorem 1* We claim that the dimension of $E^\pi$ can be bounded from above by the dimension of $E^{\pi_U}$:

Claim: For a generic distribution $\pi$, the dimension of $E^\pi$ is less than or equal to the dimension $E^{\pi_0}$ for a particular distribution $\pi_0$.

Proof of Claim: We think first of the coordinates of $\pi$ as parameters, so that we consider model invariants as linear polynomials in the variables $x_\chi$ with coefficients in the field of rational functions $\mathbb{R}(\pi)$ (i.e. the field of fractions of the ring of polynomials $\mathbb{R}[\pi_1, \ldots, \pi_\kappa]$). The set of all model invariants is a $\mathbb{R}(\pi_1, \ldots, \pi_\kappa)$-vector space. Consider a basis $l_1, \ldots, l_t$ of this space and let $E$ be its orthogonal subspace, $E = \{x \in \mathbb{R}^N | l_i(x) = 0, i = 1, \ldots, t\}$ so that $\dim E = N - t$. When we substitute $\pi$ by a particular value $\pi_0$, $l_1, \ldots, l_t$ may not be linearly independent any more, and the corresponding space $E^{\pi_0}$ may have dimension $\geq \dim E$. But for a generic $\pi$, the dimension of the corresponding space coincides with dimension of $E$ (because $\pi$ moves in an irreducible space). Therefore, for a generic $\pi$ we have $\dim E^\pi = \dim E \leq \dim E^{\pi_0}$ and the claim is proved. By the Claim, for a generic $\pi$, the dimension of $E^\pi$ is less than or equal to $\dim E^{\pi_U}$ for the uniform distribution $\pi_U$ and the dimension of this vector space is $|\Sigma_\kappa|$ (by Lemma 7(b)). Thus, $\dim E^\pi \leq |\Sigma_\kappa|$. On the other hand, the dimension of $\langle q_\sigma \mid \sigma \in \Sigma_\kappa \rangle$ is $|\Sigma_\kappa| - 1$ by Lemma 7(a). The inclusion

$$\langle q_\sigma \mid \sigma \in \Sigma_\kappa \rangle \subseteq \mathcal{D}^\pi = E^\pi \cap H$$

finishes the proof. Note that if $\kappa \geq n$ one immediately has $\mathcal{D}^\pi = \langle q_\sigma \mid \sigma \in \Sigma_n$ for any $\pi$, and its dimension follows from Lemma 7(a). □

*Remark 3* In Theorem 1 we give a set of affine independent points that span $\mathcal{D}^\pi$ for almost any distribution $\pi$. From this set of points (vectors) it easy to compute a basis of the space of linear invariants $L^\pi$ as its orthogonal space.

*Example 2* We give here the coordinates of the points that span the spaces of mixtures on trees with $n = 4$ and $\kappa = 4$ or $\kappa = 3$.

For $\kappa = 4$ we have $|\Sigma_4| = B_4 = 15$ and $\mathcal{D}^\pi = \langle q_\sigma \mid \sigma \in \Sigma_\kappa \rangle$. We start with 12 partitions $\sigma$ that correspond to forests in the star tree $T_*$. We call $q_\bullet$ the point corresponding to the trivial subforest of $T_*$ (formed by singletons). We call $q_{ij}$ the points corresponding to the full subforest of $T_*$ formed by the tree $T[i, j]$ and singletons (this gives six points, $q_{ij}, i < j$). Then we consider the forests formed by a

**Table 1** Linearly independent points for $\mathcal{D}_{T_*}$ for $n = 4$ in coordinates $\tilde{x}'s$

| | xxxx | xxxy | xxyx | xyxx | yxxx | xxyy | xyxy | xyyx | xxyz | xyxz | xyzx | yxxz | yxzx | yzxx | xyzw |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $q_\bullet$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $q_{12}$ | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{13}$ | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | 0 | 0 | 0 |
| $q_{14}$ | $\frac{1}{\pi_x}$ | 0 | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | 0 | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | 0 | 0 |
| $q_{23}$ | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_y}$ | 0 | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 | 0 |
| $q_{24}$ | $\frac{1}{\pi_x}$ | 0 | $\frac{1}{\pi_x}$ | 0 | $\frac{1}{\pi_x}$ | 0 | $\frac{1}{\pi_y}$ | 0 | 0 | 0 | 0 | 0 | $\frac{1}{\pi_x}$ | 0 | 0 |
| $q_{34}$ | $\frac{1}{\pi_x}$ | 0 | 0 | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_x}$ | $\frac{1}{\pi_y}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\frac{1}{\pi_x}$ | 0 |
| $q_{123}$ | $\frac{1}{\pi_x^2}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{124}$ | $\frac{1}{\pi_x^2}$ | 0 | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{134}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{234}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{1234}$ | $\frac{1}{\pi_x^3}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 2** The new point added for tree 12|34

| | xxxx | xxxy | xxyx | xyxx | yxxx | xxyy | xyxy | xyyx | xxyz | xyxz | xyzx | yxxz | yxzx | yzxx | xyzw |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $q_{12\|34}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | $\frac{1}{\pi_x \pi_y}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

subtree of three leaves $i$, $j$, $k$ and a singleton, which gives four points $q_{123}, q_{124}, q_{134}, q_{234}$. Finally, we denote by $q_{1234}$ the point corresponding to the forest $F = \{T_*\}$. To simplify notation we write the normalized coordinates $\tilde{x}_{\chi_1 \dots \chi_4}$ instead of $x_{\chi_1 \dots \chi_4}$. Let the space of states $S$ be $\{x, y, z, w\}$. In order to prove that the 15 points we provide are affine linearly independent, it is enough to look at the following 15 coordinates of these points:

$$\tilde{x}_{xxxx}, \tilde{x}_{xxxy}, \tilde{x}_{xxyx}, \tilde{x}_{xyxx}, \tilde{x}_{yxxx}, \tilde{x}_{xxyy}, \tilde{x}_{xyxy}, \tilde{x}_{xyyx},$$
$$\tilde{x}_{xxyz}, \tilde{x}_{xyxz}, \tilde{x}_{xyzx}, \tilde{x}_{yxzx}, \tilde{x}_{yxxz}, \tilde{x}_{yzxx}, \tilde{x}_{xyzw}.$$

In Table 1 we write the coordinates of the first 12 points considered above.

If we consider the previous points plus the point $q_{12|34}$ that corresponds to the forest $\{T[1, 2], T[3, 4]\}$ on the tree $T_{12|34}$, then we obtain a set of linearly independent points that span $\mathcal{D}_{12|34}^\pi$. In Table 2 we show the coordinates of this new point.

Now we consider the points corresponding to the forests compatible for the remaining quartets, $q_{13|24}, q_{14|23}$ (their coordinates are shown in Table 3). The previous points together with these two points span the space of mixtures $\mathcal{D}^\pi$.

**Table 3** The two points added when considering the quartets 13|24 and 14|23

| | xxxx | xxxy | xxyx | xyxx | yxxx | xxyy | xyxy | xyyx | xxyz | xyxz | xyzx | yxxz | yxzx | yzxx | xyzw |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $q_{13|24}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | $\frac{1}{\pi_x \pi_y}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $q_{14|23}$ | $\frac{1}{\pi_x^2}$ | 0 | 0 | 0 | 0 | 0 | 0 | $\frac{1}{\pi_x \pi_y}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Consider now the case $\kappa = 3$. Then, according to Theorem 1, $\mathcal{D}^\pi$ has dimension 13 for generic $\pi$. Indeed, if we consider the 15 points above, then they are no longer linearly independent when the last column of the table is removed. The last 14 points suffice to span $\mathcal{D}^\pi$ in this case.

## 6 Phylogenetic mixtures on a fixed tree

In this section we compute the dimension of the space of phylogenetic mixtures on a tree, give an algorithm to compute a basis of the space of liner topology invariants and we explain whether Lake-type invariants of Proposition 3 suffice to describe the space of phylogenetic invariants. For $\kappa = 2$ there are known to be no linear topology invariants (Matsen et al. 2008); these arise for $\kappa \geq 3$ (see Lemma 10 below, though Lake-type invariants only appear when $\kappa \geq 4$). Moreover, even when $\kappa = 4$ for certain models there exist other linear topology invariants beyond the Lake-type ones (Fu 1995). By considering the $EI/RC$ model we show how it is possible to characterize the quotient space of linear topology invariants for any number of states and taxa, and provide an explicit algorithm for constructing a basis for the (quotient) space of topological invariants. As explained in the introduction, linear topology invariants are of interest because they provide a way to distinguish distributions coming from mixtures on a particular topology from distributions arising as mixtures on another topology.

Recall that $E_T^\pi$ is the space of vectors where the linear phylogenetic invariants vanish. We know by Lemma 3(b) that a homogeneous linear polynomial vanishes on all distributions $\mathbb{P}_{T,\Theta}$ if and only if it vanishes on all distributions of type $\mathbb{P}_{T,\Theta_F}$ for $F$ a full subforest of $T$. Therefore we have

$$E_T^\pi = \langle \mathbf{q}_F | \ F \in \mathcal{F}_T \rangle.$$

*Example 3* Let $n = 3$, let $T$ be the tripod tree and assume that $\kappa \geq 3$. We prove here that the vectors $\mathbf{q}_F$, for $F \in \mathcal{F}_T$ are linearly independent. These vectors are: $\mathbf{q}_\bullet$ corresponding to the trivial subforest, $\mathbf{q}_{12|3}, \mathbf{q}_{13|2}, \mathbf{q}_{23|1}$ corresponding to full sub forests with one singleton, and $\mathbf{q}_{123}$ corresponding to the tree itself. We choose three states $x$, $y$, $z$ and we provide in Table 4 the submatrix corresponding to the coordinates $x_{xxx}, x_{xxy}, x_{xxy}, x_{xyx}, x_{yxx}, x_{xyz}$. It is clear that this submatrix has nonvanishing determinant if $\pi$ is positive.

Let $T$ be a binary tree on $[n]$, $n \geq 4$, and assume that leaves $n$ and $n - 1$ form a cherry $c$. Let $u$ be the interior node of this cherry, and let $e$ be the edge adjacent to $u$

**Table 4** Table of Example 3

| | $x_{xxx}$ | $x_{xxy}$ | $x_{xxy}$ | $x_{xyx}$ | $x_{yxx}$ | $x_{xyz}$ |
|---|---|---|---|---|---|---|
| $\mathbf{q}_\bullet$ | $\pi_x^3$ | $\pi_x^2\pi_y$ | $\pi_x^2\pi_y$ | $\pi_x^2\pi_y$ | $\pi_x^2\pi_y$ | $\pi_x\pi_y\pi_z$ |
| $\mathbf{q}_{12\mid3}$ | $\pi_x^2$ | $\pi_x\pi_y$ | $0$ | $0$ | $0$ | $0$ |
| $\mathbf{q}_{13\mid2}$ | $\pi_x^2$ | $0$ | $\pi_x\pi_y$ | $0$ | $0$ | $0$ |
| $\mathbf{q}_{23\mid1}$ | $\pi_x^2$ | $0$ | $0$ | $\pi_x\pi_y$ | $0$ | $0$ |
| $\mathbf{q}_{123}$ | $\pi_x$ | $0$ | $0$ | $0$ | $0$ | $0$ |

and not to $n, n - 1$. Let $T'$ be the subtree $T - \{e_n, e_{n-1}\}$. We denote by $\mathcal{F}_c$ the set of full subforests of $T$ that contain a tree with the cherry $c = \{e_n, e_{n-1}\}$. For any leaf $l$ we let $\mathcal{F}_l$ be the set of full subforests of $T$ that contain $l$ as a singleton and we call $T_l$ the tree obtained by replacing the two edges adjacent to $e_l$ by a single edge. Then $\mathcal{F}_T$ is the disjoint union of $\mathcal{F}_c$ and $\mathcal{F}_{n-1} \cup \mathcal{F}_n$.

**Lemma 8** *For a binary tree on $n \geq 4$ leaves we have isomorphisms of vector spaces:*

$$\langle \mathbf{q}_F \mid F \in \mathcal{F}_l \rangle \cong \langle \mathbf{q}_G \mid G \in \mathcal{F}_{T_l} \rangle, \quad \langle \mathbf{q}_F \mid F \in \mathcal{F}_c \rangle \cong \langle \mathbf{q}_G \mid G \in \mathcal{F}_{T'} \rangle.$$

*Proof* We start with the first isomorphism. For simplicity we assume $l = n$ (and for this isomorphism $n$ is not necessarily a leaf in a cherry). Let $V_n$ be the vector space $\langle \mathbf{q}_F \mid F \in \mathcal{F}_n \rangle$. For any state $s \in S$ we denote by $f^s$ the projection map from $\mathbb{R}^{\kappa^n}$ to the subspace $R_s$ corresponding to coordinates $x_{\chi_1\dots\chi_{n-1}s}$, so that we can view $\mathbb{R}^{\kappa^n}$ as the direct sum $R_{s_1} \oplus \cdots \oplus R_{s_\kappa}$. For a vector $v \in \mathbb{R}^{\kappa^n}$ we denote by $(f^{s_1}(v), \dots, f^{s_\kappa}(v))$ the decomposition of $v$ according to this direct sum. Note that if $F \in \mathcal{F}_n$, then $\mathbb{P}_T(\chi_1 \dots \chi_n \mid \Theta_F) = \pi_{\chi_n} \mathbb{P}_{T'}(\chi_1 \dots \chi_{n-1} \mid \Theta_{F\mid T_n})$. In particular, we have $f^s(\mathbf{q}_F) = \pi_s \mathbf{q}_{F\mid T'}$ for any $s \in S$ and $\mathbf{q}_F = (\pi_{s_1}\mathbf{q}_{F\mid T_n}, \dots, \pi_{s_\kappa}\mathbf{q}_{F\mid T_n})$.

We prove here that (for any $s \in S$) the linear map $f^s$ is an isomorphism between $V_n$ and the target vector space. First of all, the linear map $f^s_{\mid V_n}$ is injective. Indeed, if $f^s_{\mid V_n}(v) = 0$ for a certain $v = \sum_{F\in\mathcal{F}_n} \lambda_F \mathbf{q}_F$, then $0 = \sum_{F\in\mathcal{F}_n} \lambda_F f^s(\mathbf{q}_F) = \sum_{F\in\mathcal{F}_n} \lambda_F \pi_s \mathbf{q}_{F\mid T_n}$ and hence (assuming $\pi_s \neq 0$) $\sum_{F\in\mathcal{F}_n} \lambda_F \mathbf{q}_{F\mid T'} = 0$. This implies that $v = (0, \dots, 0)$ in $R_{s_1} \oplus \cdots \oplus R_{s_\kappa}$ and so $f^s_{\mid V_n}$ is an injective linear map.

We prove that the image of $f^s_{\mid V_n}$ is $\langle \mathbf{q}_G \mid G \in \mathcal{F}_{T_n} \rangle$. From the above, one can easily see that $\operatorname{Im} f^s_{\mid V_n}$ is contained in $\langle \mathbf{q}_G \mid G \in \mathcal{F}_{T_n} \rangle$. Now for any $G \in \mathcal{F}_{T_n}$ we shall find $\tilde{G} \in \mathcal{F}_T$ such that $\tilde{G}_{\mid T_n} = G$. If $n$ does not belong to a cherry, we consider $\tilde{G}$ to be the full subforest of $T$ defined by the singleton $\{n\}$, and the trees in $G$ (thinking of $T_n$ as a subtree of $T$). If $n$ belongs to a cherry, we can think of $T_n$ as the tree $T'$ described above. Now for any $G \in \mathcal{F}_{T'}$, we consider $\tilde{G}$ the full subforest of $T$ defined by: the singleton $\{n\}$, $t$ for any $t \in G$ not containing $e$ nor $u$, $t \cup e_{n-1}$ if there is $t \in G$ containing $e$, and the singleton $\{n - 1\}$ if $G$ contains the singleton $\{u\}$. In this way we have $\tilde{G}_{\mid T'} = G$ and $\mathbf{q}_G = \frac{1}{\pi_s} f^s_{\mid V_n} \mathbf{q}_{\tilde{G}} \in \operatorname{Im} f^s_{\mid V_n}$, so the other inclusion is proved.

As far as the second isomorphism is concerned, we consider the subspace $L \subset \mathbb{R}^{\kappa^n}$ given by coordinates of type $x_{\chi_1\dots\chi_{n-2}ss}$ for any $\chi_1, \dots, \chi_{n-2}, s$ in $S$. We have $\mathbb{R}^{\kappa^n} = L \oplus L^\perp$ and if $f$ denotes the projection to $L$, then any vector $v$ can be decomposed as $(f(v), v - f(v))$. If $F \in \mathcal{F}_c$, then $\mathbb{P}_T(\chi_1 \dots \chi_{n-1}\chi_n \mid \Theta_F)$ is zero if $\chi_{n-1} \neq \chi_n$ and is

equal to $\mathbb{P}_{T'}(\chi_1 \ldots \chi_{n-2}|\Theta_{F|T'})$ if $\chi_{n-1} = \chi_n = s$. Hence, if $F \in \mathcal{F}_c$ we have $\mathbf{q}_F = (f(\mathbf{q}), 0) = (\mathbf{q}_{F|T'}, 0)$. Now we prove that $f_{|V_c}$ is injective. Let $v = \sum_{F \in \mathcal{F}_c} \lambda_F \mathbf{q}_F$ and suppose that $f(v) = 0$. Then $0 = \sum_{F \in \mathcal{F}_c} \lambda_F f(\mathbf{q}_F) = \sum_{F \in \mathcal{F}_c} \lambda_F \mathbf{q}_{F|T'}$ and

$$
v = \sum_{F \in \mathcal{F}_c} \lambda_F \mathbf{q}_F = \sum_{F \in \mathcal{F}_c} \lambda_F (\mathbf{q}_{F|T'}, 0) = \left( \sum_{F \in \mathcal{F}_c} \lambda_F \mathbf{q}_{F|T'}, 0 \right) = 0.
$$

This proves that $f_{|V_c}$ is injective. Moreover the image of this map is included in the subspace $\langle \mathbf{q}_G | \ G \in \mathcal{F}_{T'} \rangle$. For any $G \in \mathcal{F}_{T'}$ we consider the full subforest $\bar{G}$ of $T$ defined by: the trees in $G$ that do not contain $e$, $t \cup c$ if $t$ contains $e$, and the cherry $c$ if $G$ contains the singleton $\{u\}$. Therefore we have $\bar{G}_{|T'} = G$ and $\mathbf{q}_G = f_{|V_c} \mathbf{q}_{\bar{G}} \in \mathrm{Im} f_{|V_c}^s$.

**Theorem 2** *Let $T$ a phylogenetic tree on $n$ leaves, $n \geq 3$, evolving under the $EI/RC$ model for any distribution $\pi$ on $\kappa \geq 3$ states. Then, $\{q_F | \ F \in \mathcal{F}_T\}$ are affine independent points that span the space of phylogenetic mixtures on $T$, $\mathcal{D}_T^\pi$. In particular, the dimension of $\mathcal{D}_T^\pi$ is $|\mathcal{F}_T| - 1$ and when $T$ is binary this dimension is equal to the Fibonacci number $F_{2n-1}$ minus 1.*

*Proof* We proceed by induction on $n$. The statement of the theorem is equivalent to $\dim E_T^\pi = |\mathcal{F}_T|$.

The cases $n = 3$ and $n = 4$ are handled by Examples 2 and 3.

For $n \geq 5$, suppose first that $T$ is a binary tree. We may assume that the statement is true for trees with strictly less than $n$ leaves. We suppose that $n$ and $n - 1$ form a cherry and adopt the notation fixed above. Then we have that

$$
E_T^\pi = \langle \mathbf{q}_F | \ F \in \mathcal{F}_T \rangle = \langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle + \langle \mathbf{q}_F | \ F \in \mathcal{F}_c \rangle.
$$

Note that $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle$ equals $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \rangle + \langle \mathbf{q}_F | \ F \in \mathcal{F}_n \rangle$. We know that $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \rangle$ and $\langle \mathbf{q}_F | \ F \in \mathcal{F}_n \rangle$ have dimension $|\mathcal{F}_{T'}|$ by Lemma 8 and the induction hypothesis. These subspaces intersect in $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \cap \mathcal{F}_n \rangle$. By Lemma 8 (applied twice) and the induction hypothesis, this linear space has dimension $|\mathcal{F}_{T''}|$ where $T''$ is a tree on $n - 2$ leaves. Therefore, using Grassmann's formula $(\dim(U + W) = \dim U + \dim W - \dim(U \cap W)$ for subspaces $U, W$ of a vector space) we have that $\dim(\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \rangle + \langle \mathbf{q}_F | \ F \in \mathcal{F}_n \rangle) = |\mathcal{F}_{T'}| + |\mathcal{F}_{T'}| - |\mathcal{F}_{T''}|$. As all of these trees are binary, this dimension equals the Fibonacci number $F_{2n-2}$ since $F_{2n-2} = F_{2n-3} + F_{2n-3} - F_{2n-5}$.

On the other hand, by Lemma 8 and the induction hypothesis, $\langle \mathbf{q}_F | \ F \in \mathcal{F}_c \rangle$ has dimension $|\mathcal{F}_{T'}| = F_{2n-3}$. Let us prove now that $\langle \mathbf{q}_F | \ F \in \mathcal{F}_c \rangle$ and $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle$ only intersect in the zero vector. Let $v$ be a vector in the intersection,

$$
v = \sum_{F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n} \lambda_F \mathbf{q}_F = \sum_{G \in F_c} \mu_G \mathbf{q}_G.
$$

Looking at the right-hand side we see that all the coordinates of $v$ of type $x_{\chi_1 \ldots \chi_{n-2} s s'}$ for $s \neq s'$ are zero. Let us fix $\chi_1, \ldots, \chi_{n-2}, s \in S$ and we shall prove that the coordinate

$x_{\chi_1 \ldots \chi_{n-2}ss}$ of $v$, $x_{\chi_1 \ldots \chi_{n-2}ss}(v)$, is $0$. Let us split the sum $\sum_{F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n}$ into two terms (although this decomposition may not be unique): $\sum_{F \in \mathcal{F}_{n-1}} \lambda_F \mathbf{q}_F + \sum_{H \in \mathcal{F}_n} \lambda_H \mathbf{q}_H$. We denote by $F'$ the restriction of a forest $F$ to $T'$. Note that

$$x_{\chi_1 \ldots \chi_{n-2}ss}(v) = \pi_s x_{\chi_1 \ldots \chi_{n-2}s} \left( \sum_{F \in \mathcal{F}_{n-1}} \lambda_F \mathbf{q}_{F'} \right) + \pi_s x_{\chi_1 \ldots \chi_{n-2}s} \left( \sum_{H \in \mathcal{F}_n} \lambda_H \mathbf{q}_{H'} \right).$$

For each $\alpha \in S$ we denote by $a(\alpha)$ the value of the coordinate $x_{\chi_1 \ldots \chi_{n-2}\alpha}$ of $\sum_{F \in \mathcal{F}_{n-1}} \lambda_F \mathbf{q}_{F'}$ and by $b(\alpha)$ the value of this coordinate at $\sum_{H \in \mathcal{F}_n} \lambda_H \mathbf{q}_{H'}$. We want to prove that $a(s) + b(s) = 0$. Consider $s'$ and $s''$ states in $S$ different from $s$ (this is possible because $\kappa \geq 3$). As

$$0 = x_{\chi_1 \ldots \chi_{n-2}ss'}(v) = \pi_{s'} a(s) + \pi_s b(s'),$$
$$0 = x_{\chi_1 \ldots \chi_{n-2}s's}(v) = \pi_s a(s') + \pi_{s'} b(s),$$
$$0 = x_{\chi_1 \ldots \chi_{n-2}s's''}(v) = \pi_{s''} a(s') + \pi_{s'} b(s''), \quad \text{and}$$
$$0 = x_{\chi_1 \ldots \chi_{n-2}s''s'}(v) = \pi_{s'} a(s'') + \pi_{s''} b(s'),$$

we have

$$a(s) + b(s) = -\frac{\pi_s}{\pi_{s'}} (b(s') + a(s')) = \frac{\pi'_s}{\pi_{s''}} \frac{\pi_s}{\pi_{s'}} (a(s'') + b(s'')).$$

But now we use the analogous relations between $a(s)$, $a(s'')$, $b(s)$, $b(s'')$:

$$0 = x_{\chi_1 \ldots \chi_{n-2}ss''}(v) = \pi_{s''} a(s) + \pi_s b(s'') \quad \text{and}$$
$$0 = x_{\chi_1 \ldots \chi_{n-2}s''s}(v) = \pi_s a(s'') + \pi_{s''} b(s),$$

in order to obtain that $a(s) + b(s) = -\frac{\pi_s}{\pi_{s''}} (b(s'') + a(s''))$. Therefore, $a(s) + b(s) = -a(s) - b(s)$ and this quantity vanishes.

Applying Grassmann's formula again, we have $\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle \cap \langle \mathbf{q}_F | \ F \in \mathcal{F}_c \rangle = 0$ and

$$\dim E_T^\pi = \dim(\langle \mathbf{q}_F | \ F \in \mathcal{F}_{n-1} \rangle + \langle \mathbf{q}_F | \ F \in \mathcal{F}_n \rangle) + \dim \langle \mathbf{q}_F | \ F \in \mathcal{F}_c \rangle.$$

We have already seen that the first term is equal to $F_{2n-2}$. The second term is equal to $F_{2n-3}$ by Lemma 8 and the induction hypothesis. Therefore $\dim E_T^\pi = F_{2n-1} = |\mathcal{F}_T|$.

Let us assume now that $T$ is not binary. We already know that $E_T^\pi = \langle \mathbf{q}_F | \ F \in \mathcal{F}_T \rangle$ and we only need to check that the vectors $\mathbf{q}_F$, $F \in \mathcal{F}_T$, are linearly independent. As the forests in $T$ are also subforests of any binary tree that refines $T$, these vectors are linearly independent by the binary tree case proved above. This finishes the proof.

Recall that $L^\pi = (E^\pi)^\perp$ and $L_T^\pi = (E_T^\pi)^\perp$ and therefore the quotient space $L_T^\pi / L^\pi$ of linear *topology* invariants is isomorphic to $E^\pi / E_T^\pi$. As an immediate consequence of Theorems 1 and 2 we have:

**Corollary 1** *The dimension of the space of linear topology invariants is $|\Sigma_k| - |co(T)|$ if $\pi$ is either a generic distribution or the uniform distribution, or $\kappa \geq n$ (and in this last case the dimension equals $|Inc(T)|$).*

As a consequence of Theorem 2, we are able to provide an algorithm to obtain a basis of the space of linear topology invariants for any tree $T$, $L_T^\pi / L^\pi$. To do so, note that if $proj$ is the orthogonal projection from $E^\pi$ to the subspace $L_T^\pi = (E_T^\pi)^\perp$, then $proj$ provides an isomorphism between $E^\pi / E_T^\pi$ and $L_T^\pi / L^\pi$ and therefore we have:

**Algorithm.**

1. For each $F \in \mathcal{F}_T$ compute the coordinates of the vector $\mathbf{q}_F \in E_T^\pi$.
2. Complete the basis $\{\mathbf{q}_F | \ F \in \mathcal{F}_T\}$ by vectors $v_1, \ldots, v_d$ from $E^\pi$ in order to obtain a basis of $E^\pi$.
3. Then the classes of $proj(v_1), \ldots, proj(v_d)$ form a basis of the space of linear topology invariants $L_T^\pi / L^\pi$.

Note that step 2 can be done using the Steinitz exchange lemma and the spanning set of vectors of $E^\pi$ provided in Theorem 1.

We prove now that Lake-type invariants suffice to define the space of linear topology invariants of a tree when $\kappa \geq n$ and $\pi$ is the uniform distribution. We first need a combinatorial lemma.

**Lemma 9** *For any phylogenetic tree $T$ on $[n]$ and any partition $\sigma$ that is incompatible with $T$ there exist two blocks $B$, $B'$ of $\sigma$ and leaves $x \in B$, $x' \in B'$ and an interior vertex $v$ of $T$ in the path connecting $x$ and $x'$ for which the following holds:*

> *For each leaf $l$ of $T$ in the same connected component of $T - v$ as $x$, $l \in B$ or $\{l\} \in \sigma$.*
> *For each leaf $l$ of $T$ in the same connected component of $T - v$ as $x'$, $l \in B'$ or $\{l\} \in \sigma$.*

*Proof* First suppose that $\sigma$ has no singleton blocks. Let us say that an edge $e = \{u, v\}$ of $T$ is *terminating* if:

(i) all the leaves of $T$ that are in the subtree $t_e$ of $T - v$ containing $u$ are contained in a single block of $\sigma$ (say, $B_i$), and
(ii) at least two of the other subtrees of $T - v$ contain elements of $[n]$ not in $B_i$.

For each such terminating edge $e$ delete the pendant subtree $t_e$ from $T$ and label $u$ by $B_i$. Let $T'$ be the resulting tree. This tree $T'$ has at least four leaves (since $\sigma$ is incompatible with $T$) and so $T'$ has a cherry (two leaves that are adjacent to a shared vertex $v$). This vertex $v$ and the label sets of the incident leaves ($B$ and $B'$) then satisfies the property claimed in the lemma. The extension to allow $\sigma$ to have singleton blocks is now straightforward—we can simply delete them first, repeat the argument above, and add them in afterwards.

**Corollary 2** *If $\pi_U$ is the uniform distribution and $\kappa \geq n$, then the Lake-type invariants of Proposition 3 and model invariants generate the space of linear phylogenetic invariants for $T$.*

*Proof* We omit the superscript $\pi_U$ for the spaces of linear invariants in this proof. By Lemma 7(b) the space of model invariants $L$ is spanned by the polynomials $x_\chi - x_{\chi'}$ for $\sigma(\chi) = \sigma(\chi')$ and has dimension $\kappa^n - |\Sigma_n|$ (because $\kappa \geq n$). We also have that dim $L_T = \kappa^n - \dim E_T^{\pi_U} = \kappa^n - |\mathcal{F}_T| = \kappa^n - (|\Sigma_n| - |\text{Inc}(T)|)$ and dim $L = \kappa^n - \dim E^{\pi_U} = \kappa^n - |\Sigma_n|$. Hence, we have dim $L_T/L = \dim L_T - \dim L = |\text{Inc}(T)|$. So we need to prove that Lake's invariants give a set of $|\text{Inc}(T)|$ linearly independent vectors in $L_T/L$.

Note that in $L_T/L$ we can work with polynomials in indeterminates $x_\sigma$, $\sigma \in \Sigma_n$.

Let us prove that, if $\sigma$ is an incompatible partition on $T$, then $x_\sigma$ is a linear combination of $x_{\sigma'}$ for compatible partitions $\sigma'$ of size $> |\sigma|$. To this end, we proceed by induction on $m = n - |\sigma|$.

If $m = 0$ or $1$, then $\sigma$ is convex on $T$ and there is nothing to prove. Let $m \geq 2$ and assume that we have proved the statement when $n - |\sigma|$ is smaller than $m$. Let $\sigma = \{B_1, \ldots, B_r\}$ and we call $s_1, \ldots, s_r$ the states associated to $\sigma$. Assume first that $\sigma$ has no singletons. Then, according to Lemma 9 we can find two blocks of $\sigma$, say $B_1$, $B_2$, and an interior vertex $v$ for which all leaves in one of the subtrees $T_1'$ of $T - v$ are in $B_1$, and all leaves in one of the other subtrees $T_2'$ of $T - v$ are in $B_2$. We write $l_i'$ for the set of leaves in $T_i'$ so that $B_i$ is the disjoint union of $l_i'$ and another set $l_i$. We let $\mathcal{E}$ be the event that leaves $B_i$ are in state $s_i$ for $i \geq 3$, leaves in $l_1$ are in state $s_1$ and leaves in $l_2$ are in state $s_2$. As the fully symmetric model satisfies the partial separability property (PS) and as $|\sigma| \leq n - 2 \leq \kappa - 2$, we can consider two new states $s_1'$, $s_2'$ to apply Proposition 3 (with $t = T_1'$ and $t' = T_2'$). Thus we obtain the following linear invariant (written in terms of partitions because the states do not matter, as soon as they are different):

$$x_\sigma + x_{l_1|l_1'|l_2|l_2'|B_3|\ldots|B_r} - x_{l_1|l_1'|B_2|B_3|\ldots|B_r} - x_{B_1|l_2|l_2'|B_3|\ldots|B_r}.$$

Note that all partitions involved in this expression, except for $\sigma$, have size larger than $|\sigma|$ and we can apply the induction hypothesis to any $x_{\sigma'}$ appearing here with $\sigma'$ incompatible, to write $x_\sigma$ as a linear combination of $x_{\sigma'}$'s using only compatible $\sigma'$.

If $\sigma$ has singletons, we remove these singletons in $T$ and $\sigma$ obtaining a tree $T_0$ and a partition $\sigma_0$ without singletons on $T_0$. We apply the previous argument to $\sigma_0$ and $T_0$ to obtain a linear invariant. Then we apply the Extension Lemma 6(a) recursively to add singletons and we end up also with a linear polynomial that involves $\sigma$ and partitions of larger size. Hence, we can apply the induction hypothesis again.

The linear invariants obtained in this way for each incompatible partition $\sigma$ are of Lake-type and form a set of linearly independent vectors in $L_T/L$ because they involve partitions of larger size.

*Remark 4* **Case $\kappa = 2$.** For $\kappa = 2$, Theorem 2 and Corollary 2 do not apply. In this case it is already known (see Matsen et al. 2008) that there are no linear topology invariants for the uniform distribution $\pi_U$ and hence $\mathcal{D}_T^{\pi_U} = \mathcal{D}^{\pi_U}$ for any tree $T$ (see Matsen et al. 2008). One can actually prove that this also holds for any generic distribution $\pi$ and this space has dimension $|\Sigma_2| = 2^{n-1} - 1$, see Matsen et al. (2008).

*Remark 5* **Case $\kappa = 3$.** For $\kappa = 3$ and $n = 4$, we cannot apply Corollary 2 either. But in this case we can provide another topology invariant. We describe it in the following

lemma for $n = 4$ but can be easily generalized for the uniform distribution to any tree by using a similar argument as in Proposition 3. Moreover, it is not difficult to see that for $\kappa \geq 4$ it can be derived form Lake-type invariants.

**Lemma 10** *For the tree* $12|34$ *and any positive distribution* $\pi$ *on a set $S$ of $\kappa \geq 3$ states, the polynomial*

$$\tilde{x}_{xyxy} + \tilde{x}_{xyyz} + \tilde{x}_{xyzx} - \tilde{x}_{xyyx} - \tilde{x}_{xyxz} - \tilde{x}_{xyzy}, \tag{6}$$

*for any three different states $x, y, z \in S$, is a topology invariant if $T$ evolves under the EI/RC model.*

*Proof* According to Lemma 3 we need to prove that (6) vanishes when we evaluate it at the points $q_F$, $F \in \mathcal{F}_T$. If $F$ is a forest such that $\sigma(F)$ does not refine any of the partitions $\{\{1, 3\}, \{2, 4\}\}, \{\{1, 4\}, \{2, 3\}\}$, then the coordinates that appear in (6) are all zero. If $\sigma(F)$ refines $\{\{1, 3\}, \{2, 4\}\}$, then $\sigma(F)$ is either $\{\{1, 3\}, \{2\}, \{4\}\}$, or $\{\{2, 4\}, \{1\}, \{3\}\}$ or the trivial forest. In the first two cases (6) evaluated at $q_F$ vanishes. As the evaluation of any coordinate $\tilde{x}$ at the point associated to the trivial forest is one, it also vanishes in this case. The remaining cases follow from the symmetry of leaves 3 and 4 in (6).

*Remark 6* **Case** $\kappa = 4$. For $n = 5$ not all linear topology invariants are of Lake-type. In Fu (1995) a complete list of $17 (= |\Sigma_4| - |co(T)| = 61 - 34)$ linear invariants that generate the space of linear topology invariants is given. For example, for the fully symmetric model on the set of states $\{x, y, z, w\}$ (i.e. Jukes–Cantor model),

$$x_{xyyxy} + x_{xyzwz} - x_{xyyzy} - x_{xyzxz}$$

is a topology linear invariant that cannot be described by Proposition 3.

## 7 Explicit linear invariants for quartet trees

In this section we assume that $\kappa \geq 4$ and we shall deal with quartet trees and the star tree on four leaves. Note that in the previous section we gave an explicit description of linear phylogenetic invariants only when the distribution was uniform. For a generic distribution $\pi$ we managed to compute the dimension of the space of linear phylogenetic invariants, but we did not provide a explicit set of generators. We do it in this section for the case $n = 4$, $\kappa \geq 4$, and any distribution $\pi$.

*Remark 7* In the case of quartet trees on the set of taxa $X = [4]$, the possible tree topologies are $12|34$, $13|24$, $14|23$, and the star tree $T_*$. As the star tree is a subtree of the others, the vector space of phylogenetic mixtures is

$$E^\pi = \langle \mathbf{q}_F \mid F \in \mathcal{F}_{12|34} \rangle + \langle \mathbf{q}_F \mid F \in \mathcal{F}_{12|34} \rangle + \langle \mathbf{q}_F \mid F \in \mathcal{F}_{13|24} \rangle + \langle \mathbf{q}_F \mid F \in \mathcal{F}_{14|23} \rangle.$$

By Theorem 2 we know that the vectors $\mathbf{q}_F$ are linearly independent if we let $F$ move in the set of full subforests of the tree $A|B$. As $\mathcal{F}_{12|34}$, $\mathcal{F}_{13|24}$ and $\mathcal{F}_{14|23}$ intersect at

the set of forests for the star tree $T_*$, in order to check whether a homogeneous linear polynomial vanishes at the vectors of $E^\pi$ one needs to check whether it vanishes at the 15 vectors of Tables 1, 2 and 3 that correspond to 12 subforests of $T_*$ and one forest $\mathbf{q}_{A|B}$ for each refined quartet).

**Proposition 4** *Let $x$, $y$, $z$, $w$ be four different states and define*

$$
\begin{aligned}
\beta_{x,y} = {} & \pi_x^2 \tilde{x}_{xxxy} + \pi_x \pi_y [\tilde{x}_{xxyy} + \tilde{x}_{xyxy} + \tilde{x}_{yxxy}] \\
& + \pi_x \pi_y [\tilde{x}_{zwxx} + \tilde{x}_{zxwx} + \tilde{x}_{xzwx}] + \pi_y^2 \tilde{x}_{xyzw}, \\
\delta_{x,y} = {} & \pi_x^2 [\pi_x \tilde{x}_{xxxx} + \pi_y \tilde{x}_{xxxy} + \pi_z \tilde{x}_{xxxz} + \pi_w \tilde{x}_{xxxw}] \\
& + \pi_x \pi_y [\pi_x \tilde{x}_{xxyx} + \pi_y \tilde{x}_{xxyy} + \pi_z \tilde{x}_{xxyz} + \pi_w \tilde{x}_{xxyw}] \\
& + \pi_x \pi_y [\pi_x \tilde{x}_{xyxx} + \pi_y \tilde{x}_{xyxy} + \pi_z \tilde{x}_{xyxz} + \pi_w \tilde{x}_{xyxw}] \\
& + \pi_x \pi_y [\pi_x \tilde{x}_{yxxx} + \pi_y \tilde{x}_{yxxy} + \pi_z \tilde{x}_{yxxz} + \pi_w \tilde{x}_{yxxw}] \\
& + \pi_y^2 [\pi_x \tilde{x}_{xyzx} + \pi_y \tilde{x}_{xyzy} + \pi_z \tilde{x}_{xyzz} + \pi_w \tilde{x}_{xyzw}].
\end{aligned}
$$

*Then following are linear model invariants for quartet trees evolving under the EI/RC model:*

$$
\pi_y \tilde{x}_{xxyy} + \pi_z \tilde{x}_{xxyz} - \pi_y \tilde{x}_{xxzy} + \pi_z \tilde{x}_{xxzz} \tag{7}
$$
$$
\pi_x \tilde{x}_{xxyz} + \pi_w \tilde{x}_{xwyz} - \pi_w \tilde{x}_{wwyz} + \pi_x \tilde{x}_{wxyz} \tag{8}
$$
$$
\beta_{x,y} - \beta_{y,x} \tag{9}
$$
$$
\delta_{x,y} - \delta_{y,x} \tag{10}
$$

*One obtains analogous linear model invariants by considering any permutation of the set of leaves.*

*Proof* From the extension Lemma 6(b) it follows that (7) and (8) are model invariants. Indeed, if we consider the star tree $T_2$ on two leaves, then it is easy to check that

$$
\pi_y \tilde{x}_{yy} + \pi_z \tilde{x}_{yz} - \pi_y \tilde{x}_{zy} - \pi_z \tilde{x}_{zz}
$$

is a linear phylogenetic invariant. By identifying $T_2$ with the star tree $T_{3,4}$ on leaves 3, 4 we can apply Lemma 6(b) with $\mu = xx$ to obtain (7) for the quartet tree $T = 12|34$ (because $T$ can be obtained by attaching the tripod tree $T_{1,2,l}$ to the edge leading to leaf 3 of $T_2$). In particular, (7) vanishes for the star tree $T_*$ on four leaves. Similarly, in order to see that (8) is a phylogenetic invariant for the star tree $T_*$, we use the phylogenetic invariant

$$
\pi_x \tilde{x}_{xx} + \pi_w \tilde{x}_{xw} - \pi_w \tilde{x}_{ww} - \pi_x \tilde{x}_{wx}
$$

for the tree $T_2 = T_{1,2}$ and apply Lemma 6(b) with $\mu = yz$. By Lemma 6(c) we see that (8) is a phylogenetic invariant for the quartet tree 12|34 (and hence also for the star tree $T_*$).

In order to prove that (7) and (8) are model invariants, it only remains to check that these expression vanish when evaluated at $\mathbf{q}_{13|24}$ and $\mathbf{q}_{14|23}$, which is straight forward because all coordinates involved in the expressions are 0 for these vectors.

We check now that (9) and (10) are model invariants having Remark 7 in mind. Looking at Table 1, we observe that $\beta_{x,y}$ (respectively $\delta_{x,y}$) evaluated at $\mathbf{q}_\bullet$ is $\pi_x^2 + 6\pi_x\pi_y + \pi_y^2$ (resp. $\pi_x^2 + 3\pi_x\pi_y + \pi_y^2(\pi_x + \pi_y + \pi_z + \pi_w)$). As these expressions are symmetric for $x$ and $y$, (9) and (10) vanish in this case.

Now we consider the other vectors in Table 1, $\mathbf{q}_B$, where $B$ is a block of $m$ leaves, $m \geq 2$, and the partition associated to this point is $B$ and singleton blocks.

We start with $m = 2$. Using the equalities of lemma 4, we can see that $\beta_{x,y}$ and $\delta_{x,y}$ are symmetric under the permutation of leaves 1,2, and 3. Thus we only need to consider that $B$ is formed either by $\{1, 2\}$ or by $\{3, 4\}$. In the first case, $\beta_{x,y}$ evaluated at $\mathbf{q}_B$ is $\pi_x + \pi_y$ and $\delta_{x,y}$ is $(\pi_x + \pi_y)(\pi_x + \pi_y + \pi_z + \pi_w)$. As these expressions are symmetric in $x$ and $y$, (9) and (10) also vanish in this case. If $B = \{3, 4\}$, then the evaluation of $\beta_{x,y}$ at $\mathbf{q}_B$ equals $\pi_x + \pi_y$ and the evaluation of (10) gives $\pi_x^2 + 3\pi_x\pi_y + \pi_y^2$. Again, these are symmetric in $x$, $y$ and (9), (10) vanish.

Now we consider $m = 3$. Let us assume first that $B = \{1, 2, 3\}$. In this case, the evaluation of $\beta_{x,y}$ at $\mathbf{q}_B$ equals 1 and the evaluation of $\delta x$, $y$ is $\pi_x + \pi_y + \pi_z + \pi_w$. Therefore (9) and (10) vanish at $\mathbf{q}_B$. If $B$ contains the leaf 4, then all terms in the evaluation of $\beta_{x,y}$ at $\mathbf{q}_B$ are zero and the evaluation of $\delta x$, $y$ at $\mathbf{q}_B$ is $\pi_x + \pi_y$. Therefore (9) and (10) also hold for these vectors.

If $m = 4$, then (9) vanishes trivially because all its terms are 0. Moreover $\delta x$, $y$ is equal to 1 when evaluated at $\mathbf{q}_{1234}$ and there fore both equations hold for this vector.

The only remaining cases to check correspond to the vectors $\mathbf{q}_{12|34}$, $\mathbf{q}_{13|24}$ and $\mathbf{q}_{14|23}$ of Tables 2 and 3. As $\beta_{x,y}$ is equal to 1 and $\delta_{x,y}$ is equal to $\pi_x + \pi_y$ when these expressions are evaluated at these vectors, both Eqs. (9) and (10) vanish on these vectors.

Note that when we apply a permutation of the set of leaves, the resulting polynomials are phylogenetic invariants because we have just proven that the original ones are linear model invariants.

**Theorem 3** *For any distribution $\pi$, the space of linear model invariants $L^\pi$ for $n = 4$ and $\kappa \geq 4$ is generated by the phylogenetic invariants of Proposition 4 together with $\tilde{x}_\chi - \tilde{x}_{\chi'}$ for any $\chi \equiv \chi'$ and has dimension $\kappa^4 - B_4 = \kappa^4 - 15$.*

For the fully symmetric model we have already seen in Remark 2 that $x_\chi - x_{\chi'}$ are linear phylogenetic invariants if $\sigma(\chi) = \sigma(\chi')$. In this case this set of invariants defines the same vector space as the phylogenetic invariants in Theorem 3.

*Remark 8* Although one could replace (9) by other phylogenetic invariants obtained from marginalization from a phylogenetic invariant relating $\tilde{x}_{xxy}$ and $\tilde{x}_{yyx}$ on the tripod, this expression would have less symmetries than (9) and therefore we decided to use (9) instead (similarly for (10)).

*Proof* We let $F^\pi$ be the space of vectors where all the linear polynomials in the statement vanish. Then we shall prove that for the vectors in $F^\pi$, any coordinate $\tilde{x}_\chi$

can be expressed as a linear combination of the following 15 coordinates:

$$\tilde{x}_{xxxx}$$
$$\tilde{x}_{xxxy}, \tilde{x}_{xxyx}, \tilde{x}_{xyxx}, \tilde{x}_{yxxx}$$
$$\tilde{x}_{xxyy}, \tilde{x}_{xyxy}, \tilde{x}_{xyyx}$$
$$\tilde{x}_{xxyz}, \tilde{x}_{xyxz}, \tilde{x}_{xyzx}, \tilde{x}_{yxzx}, \tilde{x}_{yxxz}, \tilde{x}_{yzxx}$$
$$\tilde{x}_{xyzw}$$

This will prove that $F^\pi$ is a vector space of dimension 15 or lower. By Lemma 7 we know that dim $\mathcal{D}^\pi$ is $\geq |\Sigma_\kappa| - 1$, which is $B_4 - 1 = 14$ for $n = 4$. As we have the inclusion $\mathcal{D}^\pi = E^\pi \cap H \subseteq F^\pi \cap H$ this will finish the proof.

First note that by Lemma 4 we have $\tilde{x}_{xxxy'} = \tilde{x}_{xxxy}, \tilde{x}_{xxy'z'} = \tilde{x}_{xxyz}, \tilde{x}_{x'y'z'w'} = \tilde{x}_{xyzw}$ for any $y' \neq x, x', z' \neq y, y', x, x', w' \neq x, y, z, x', y', z'$.

Using the equation (8) = 0 one can write $\tilde{x}_{x'x'y'z'}$ as a linear combination of $\tilde{x}_{xxyz}$ and $\tilde{x}_{xyzw}$. The Eq. (7) = 0 allows us to put $\tilde{x}_{xxy'y'}$ as a linear combination of $\tilde{x}_{xxyy}$ if $y' \neq y$. In order to write $\tilde{x}_{yyxx}$ (or similarly $\tilde{x}_{yxxy}$) in terms of the allowed coordinates we need to do two steps. We use expression (7) three times to put first $\tilde{x}_{yyxx}$ in terms of $\tilde{x}_{yyzz}$ first, then $\tilde{x}_{yyzz}$ in terms of $\tilde{x}_{xxzz}$ and finally $\tilde{x}_{xxzz}$ in terms of $\tilde{x}_{xxyy}$. Interchanging the role of leaves 1,2 with 3,4 we also obtain $\tilde{x}_{x'x'yy}$ as a linear combination of $\tilde{x}_{xxyy}$ if $x' \neq x$. In the same way, we can use the Eq. (9) = 0 to put $\tilde{x}_{x'x'x'y'}$ as a linear combination of $\tilde{x}_{xxxy}$ and other coordinates which we now know that are linear combinations of the allowed coordinates. Finally, we use the Eq. (10) = 0 to put $\tilde{x}_{x'x'x'x'}$ for $x' \neq x$ as a linear combination of $\tilde{x}_{xxxx}$ and other allowed coordinates.

By considering these relations above and all permutations of the leaves, we end up with every coordinate written as a linear combination of the allowed list of 15 coordinates.

We now consider the two linear topology invariants that we obtained in Example 1: in terms of the $\tilde{x}'s$ above, the corresponding equations for the quartet tree 12|34 these are

$$H_1 : \quad \tilde{x}_{xyxy} + \tilde{x}_{xyzw} = \tilde{x}_{xyzy} + \tilde{x}_{xyxw}$$
$$H_2 : \quad \tilde{x}_{xyyx} + \tilde{x}_{xywz} = \tilde{x}_{xyyz} + \tilde{x}_{xywx}.$$

Equations $H_1$ and $H_2$ are linearly independent and drop the dimension by two. In total, we have that $\mathcal{D}^\pi_{12|34}$ is contained in an affine space $E^\pi \cap H \cap H_1 \cap H_2$ of dimension 12. As the dimension of $\mathcal{D}^\pi_{12|34}$ is 12 and for the star tree dim $\mathcal{D}^\pi_{T_*} = 11$ we have:

**Corollary 3** *For $n = 4$ and any distribution $\pi$ one has*

$$\mathcal{D}^\pi = E^\pi \cap H$$
$$\mathcal{D}^\pi_{12|34} = E^\pi \cap H \cap H_1 \cap H_2$$
$$\mathcal{D}^\pi_{T_*} = E^\pi \cap H \cap H_1 \cap H_2 \cap H_3$$

*where $H_3 : \tilde{x}_{xxyy} + \tilde{x}_{xzyw} = \tilde{x}_{xzyy} + \tilde{x}_{xxyw}$ and $T_*$ denotes the star tree on four leaves. In particular, Lake-type invariants generate all linear topology invariants for quartet trees evolving under the EI model.*

## 8 The infinite-state random cluster model $RC_\infty$

Recall that in the random cluster model, each edge of $T$ is cut with some probability $\theta_e$ to obtain a resulting partition $\sigma$ of the leaf set $X$. Each block is then assigned a state independently according to the distribution $\pi$. However, we could just consider the partition $\sigma$ itself as the output of this process (rather than assigning states, which has the effect of combining some blocks together when they receive the same state). We call this the *infinite state RC model $RC_\infty$* since it has a natural interpretation as the limiting distribution on partitions induced by the $EI/RC$ model as the number of states $\kappa$ in $S$ tends to infinity when states have at least roughly similar probabilities.

More precisely, under the $RC$ model, the probability that two blocks of $\sigma$ are assigned a same state in the equal input model is at most $n \sum_{\alpha \in S} \pi_\alpha^2$, by Boole's inequality (note that there are at most $n$ blocks in $\sigma$). Suppose that $\pi_\alpha \in [a/k, b/k]$ for some fixed $a, b$ then as $k = |S| \to \infty$ all blocks of $\sigma$ receive distinct states with probability converging to 1 (this restriction on $\pi$ can be weakened a little further). The $RC_\infty$ model is sometimes referred to as the 'Kimura's infinite alleles' model in phylogenetics, and it was studied mathematically in Mossel and Steel (2004).

### 8.1 Linear invariants for $RC_\infty$

The linear phylogenetic invariants for the infinite-state random cluster model are particularly easy to describe.

Let $p_\sigma = \mathbb{P}_T(\sigma|\Theta)$ be the probability of generating partition $\sigma$ on $T$ under the $RC_\infty$ model with edge cut probabilities $\Theta = (\theta_e)$, and recall the definitions of $co(T)$ and $Inc(T)$ from Sect. 4.1.

**Proposition 5** *Under the $RC_\infty$ model:*

(i) $\mathbb{P}_T(\sigma|\Theta) = 0$ *for all $\Theta$ if and only if $\sigma \in Inc(T)$.*
(ii) $\{x_\sigma : \sigma \in Inc(T)\}$ *forms a basis for the vector space $L_T$ of linear phylogenetic invariants for $T$ and of the space of linear topology invariants. Consequently, this space has dimension $|Inc(T)| = B_n - |co(T)|$.*
(iii) *The space of all phylogenetic mixtures on $T$ has dimension $|co(T)| - 1$.*
(iv) *The space of all phylogenetic mixtures on all n-leaf trees under the $RC_\infty$ model has dimension $B_n - 1$.*

*Proof* (i) Suppose that $\sigma \in Inc(T)$. Then there exists two blocks $B, B'$ of $\sigma$ and leaves $x, y \in B$ and $x', y' \in B'$ for which the paths $P(T; x, y)$ and $P(T; x', y')$ share at least one vertex. Now since $x, y \in B$ and $x', y' \in B'$ the only way to generate $\sigma$ under $RC_\infty$ is if none of the edges in the two paths $P(T; x, y)$ and $P(T; x', y')$ is cut. Since these paths intersect on a vertex this implies that $x$ and $x'$ must be the same block, i.e. that $B = B'$. Thus $\sigma$ cannot be generated with

positive probability under the $RC_\infty$ model. Conversely, suppose that $\sigma$ is convex on $T$. Then set $\theta_e = 0$ for all edges in $\{T[B] : B \in \sigma\}$ and set $\theta_e = 1$ for all other edges. Then $p_\sigma = 1$.

(ii) If $\sum \lambda_\sigma x_\sigma$ is a linear phylogenetic invariant, then for any $\sigma$ convex on $T$ we can choose a set of parameters $\Theta$ such that $p_\sigma = 1$ (see above). This implies that $\lambda_\sigma = 0$ for any $\sigma \in \mathrm{co}(T)$. This and (i) show that the set spans the space of all linear phylogenetic invariants, and linear independence follows immediately from the observation that each polynomial involves a variable not present in any other polynomial in this set. Note that all these polynomials are topology invariants.

(iii) The space of phylogenetic mixtures $\mathcal{D}_T$ on $T$ is equal to $E_T \cap H$ where $E_T$ is the space of vectors on which the linear phylogenetic invariants vanish and $H$ is the hyperplane defined by the trivial equation $\sum_\sigma x_\sigma = 1$ (the sum is over all partitions of $[n]$). By (ii), $E_T$ has dimension $B_n - \mathrm{Inc}(T) = |\mathrm{co}(T)|$ and we are done.

(iv) Note that in the basis $\{x_\sigma : \sigma \in \mathrm{Inc}(T)\}$ of (ii) there are no model invariants. Therefore, the set $\mathcal{D}$ of phylogenetic mixtures on all trees coincides with the trivial hyperplane $H$ and has dimension $B_n - 1$.

The construction of certain quadratic phylogenetic invariants for $RC_\infty$ is also quite easy. Let $x \sim y$ denote the event that $x$ and $y$ are in the same block of the partition generated by a phylogeny under the $RC_\infty$ model, and let $p(x, y)$ denote the probability of that event. Note that $p(x, y)$ is a sum of $p_\sigma$ values over all $\sigma$ for which $x$ and $y$ are in the same block. Then $p(x, y) = \prod_{e \in P(T;x,y)} (1 - \theta_e)$, where $P(T; x, y)$ is the path in $T$ between $x$ and $y$. It follows (from the four point condition) that if the quartet tree obtained by restricting $T$ to $x, y, w, z$ is either $xy|wz$ or the star tree, then

$$p(x, w)p(y, z) - p(x, z)p(y, w) = 0.$$

## 9 Future work

It would be interesting to generalize Lake-type invariants in such a way that they generate the space of linear topology invariants for $\kappa < n$ (*cf.* Corollary 2). On the other hand, it also would be useful to give explicit linear model invariants (with many symmetries) for any number of leaves, as was done in Sect. 4 for $n = 3, 4$. These model invariants could be used for model selection as it was done in Kedzierska et al. (2012) for the uniform distribution. Extending the work of Sect. 4 to other models is also of interest because this would increase the range of models that can be considered in certain model selection software such as SPIn (http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl).

# References

Allman ES, Rhodes JA, Sullivant S (2012) When do phylogenetic mixture models mimic other phylogenetic models? Syst Biol 61:1049–1059

Casanellas M, Fernández-Sánchez J, Kedzierska AM (2012) The space of phylogenetic mixtures for equivariant models. Algorithms Mol Biol 7:33

Casanellas M, Fernández-Sánchez J (2011) Relevant phylogenetic invariants of evolutionary models. J Math Pures Appl 96:207–229

Chang JT (1996) Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci 137:51–73

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

Fernández-Sánchez J, Casanellas M (2016) Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. Syst Biol 65:280–291

Fu YX (1995) Linear invariants under Jukes' and Cantor's one-parameter model. J Theor Biol 173:339–352

Fu YX, Li WH (1991) Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree. Math Biosci 105:229–238

Kedzierska A, Drton M, Guigó R, Casanellas M (2012) SPIn: model selection for phylogenetic mixtures via linear invariants. Mol Biol Evol 29:929–937

Kemeny JG, Snell JL (1976) Finite Markov chains. Springer, New York

Lake J (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191

Matsen FA, Mossel E, Steel M (2008) Mixed-up trees: the structure of phylogenetic mixtures. Bull Math Biol 70:1115–1139

Mossel E, Steel M (2004) A phase transition for a random cluster model on phylogenetic trees. Math Biosci 187:189–203

Semple C, Steel M (2003) Phylogenetics. Oxford University Press, Oxford

Steel MA, Székely LA, Hendy MD (1994) Reconstructing trees when sequence sites evolve at variable rates. J Comput Biol 1:153–163

Steel M (2011) Can we avoid 'SIN' in the house of 'no common mechanism'? Syst Biol 60:96–109

Steel MA, Fu YX (1995) Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model. J Comput Biol 2:39–47

Štefakovič D, Vigoda E (2007) Phylogeny of mixture models: robustness of maximum likelihood and non-identifiable distributions. J Comput Biol 14:156–189

Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12:204–228