The Algebra of the General Markov Model on Phylogenetic Trees and Networks

J.G. Sumner · B.R. Holland · P.D. Jarvis

Received: 29 December 2010 / Accepted: 9 August 2011 / Published online: 6 October 2011
© Society for Mathematical Biology 2011

Abstract It is known that the Kimura 3ST model of sequence evolution on phylogenetic trees can be extended quite naturally to arbitrary split systems. However, this extension relies heavily on mathematical peculiarities of the associated Hadamard transformation, and providing an analogous augmentation of the general Markov model has thus far been elusive. In this paper, we rectify this shortcoming by showing how to extend the general Markov model on trees to include incompatible edges; and even further to more general network models. This is achieved by exploring the algebra of the generators of the continuous-time Markov chain together with the “splitting” operator that generates the branching process on phylogenetic trees. For simplicity, we proceed by discussing the two state case and then show that our results are easily extended to more states with little complication. Intriguingly, upon restriction of the two state general Markov model to the parameter space of the binary symmetric model, our extension is indistinguishable from the Hadamard approach only on trees; as soon as any incompatible splits are introduced the two approaches give rise to differing probability distributions with disparate structure. Through exploration of a simple example, we give an argument that our extension to more general networks has desirable properties that the previous approaches do not share. In particular, our construction allows for convergent evolution of previously divergent lineages; a property that is of significant interest for biological applications.

Keywords Split system · Cluster system · Markov process · Maximum likelihood

1 Introduction

Phylogenetic methods seek to infer the prior evolutionary relationships of extant, and sometimes extinct, taxa. Classically, this was achieved by comparing morphological
features, but modern methods focus on molecular data such as DNA. Harking back to sketches in Darwin’s early notebooks, it has also been assumed that evolutionary history resembles a tree structure. However, it is now well known that evolutionary processes such as hybridisation, the stochastic nature of the coalescent process on different loci, horizontal gene transfer, and recombination can all lead to the taxa under consideration being better described by a “mosaic” of trees rather than a single tree. Considering a sequence alignment of the taxa, this means that different sites within alignment may have evolved on different trees. Even when the underlying historical signal fits a tree, there may be conflicting non-historical signals caused by sampling error, long-branch attraction, nucleotide composition bias, or changes in the substitution rate at individual sites across the tree, as well as alignment or misreading errors.

It is clear that imposing a strictly treelike evolutionary history may be inappropriate in the situations described above, hence methods that can assist in identifying and understanding conflict in phylogenetic data are essential. One class of methods that have proved useful in this respect are weighted split-systems and their corresponding visualization as networks. Split networks initially arose as means of visualizing the split decomposition of a dissimilarity measure as defined by Bandelt and Dress (1992). In that work, the authors gave a decomposition that provides a way of assessing whether the structure of a distance matrix is treelike or if it contains other conflicting signals. However, the idea of a weighted split-system is very general and has arisen in many phylogenetic contexts. These include (1) median networks (Bandelt 1994), where splits and their weights are derived from a binary coding of a sequence alignment; (2) Hadamard (or spectral) analysis (Hendy and Penny 1989), which defines an invertible relationship between site patterns and a split spectrum under certain simple models (Kimura 3ST and submodels); (3) Neighbor-Net (Bryant and Moulton 2004), a distance-based method which applies a greedy agglomerative algorithm to find a circular ordering of taxa and then a least-squares approach to find weights for the corresponding set of circular splits; (4) Consensus networks (Holland and Moulton 2003), which take a set of trees and define a weighted split-system based on the number of trees that display a particular edge, possibly incorporating edge weight information (Holland et al. 2006). More recently, there has been work on developing so-called explicit network methods that describe evolutionary history as a directed acyclic graph that can include both reticulate and tree nodes (Huson et al. 2011, pp. 76–81). These methods are predominantly combinatorial; taking trees, clusters, or triples as their input.

There has been some previous work on calculating likelihood scores for particular phylogenetic networks under special models. Von Haeseler and Churchill (1993) developed a framework for computing the likelihood of a split-system for binary sequences under a Cavender–Farris model with invariable sites. Strimmer and Moulton (2000) developed a Bayesian approach that calculated the likelihood of a given directed acyclic graph (DAG) for more complex models of sequence evolution. More recently Jin et al. (2006) defined a likelihood score for phylogenetic networks as a weighted mixture of tree likelihoods.

A different approach to specifying a DAG or a mixture of trees is to model evolution with Markov models on arbitrary systems of splits or clusters. Recently, Bashford et al. (2004) and Bryant (2005, 2009) have examined the nature of the Kimura
3ST and binary-symmetric model as, under a simple extension (9) in Bashford et al. (2004); these models permit the inclusion of arbitrary splits over and above those that come from a single tree. However, these so-called “group-based” models of sequence evolution are of a very specific form and can be of limited biological validity in certain scenarios. Specifically, all group-based models have doubly-stochastic rate matrices, and thus uniform stationary distributions. This is clearly inappropriate given that varying GC content is known to be of crucial importance in phylogenetics (Jermiin et al. 2004). In fact, this was identified by Bryant (2009) as an important open question. To date, it has not been possible to employ general model-based methods to infer split networks from phylogenetic data sets.

In this article, we show how Markov models of phylogenetic evolution on trees (thought of as compatible cluster systems) can be generalized to the case of arbitrary cluster systems. In the binary case of two character states, we achieve this by studying the algebra of the generators of the continuous-time Markov chain together with the “splitting” operator that generates the branching process on phylogenetic trees. The resulting presentation of the general Markov rate matrix model on a tree is such that it can be generalized in a natural way to arbitrary cluster systems; including those that are incompatible with each other. This results in a very general model that contains the standard tree model as a special case, but has the potential to associate an individual weight and rate matrix to any additional cluster that we wish to include. Additionally, we show by example that our approach gives rise to the possibility of Markov models on much more general networks, with phylogenetic evolution proceeding in a series of “epochs” consisting of divergence or convergence of arbitrary clusters (i.e. lineages). As part of the discussion, we note that our results are fully generalizable to any number of character states with complication of detail only. Intriguingly, we will also show that under a restriction of the parameter space to the binary-symmetric case that this model does not yield the Hadamard based approach given by Bryant (2009). We close with a simple example that shows our construction has the ability to model convergent evolution of lineages; a property that is simply not available to the Hadamard based approach.

2 Preliminaries

In this article, a rooted tree is a connected directed acyclic graph $T$ with a single vertex $\rho$ (the root) of in-degree 0 and out-degree 2, $n$ vertices of in-degree 1 and out-degree 0 (the leaves), and $n-2$ vertices of in-degree 1 and out-degree 2. The leaf set is bijective with $[n] := \{1, 2, \ldots, n\}$, with each leaf labelled by $i \in [n]$. The edges of $T$ are directed away from $\rho$ and each non-leaf vertex $u$ is labelled by the set of leaves below $u$. Each directed edge $e = (u, v)$ is labelled by the label of $v$. A cluster system of $[n]$ is a set $\{A_1, A_2, \ldots, A_q\}$ where each $A_i \subset [n]$. Hence, the set of labels of the edges of $T$ forms a cluster system of $[n]$. See Fig. 1 for a graphical representation of a rooted tree as a cluster system.
Consider the vector space $V \cong \mathbb{C}^2$ with the ordered basis\(^1\)

\[
\{|0\rangle \equiv e_0 := \left( \begin{array}{c} 1 \\ 0 \end{array} \right), \quad |1\rangle \equiv e_1 := \left( \begin{array}{c} 0 \\ 1 \end{array} \right) \}
\]

With respect to this basis, we can define “Markov generators” as the zero column-sum matrices

\[
L_\alpha = \left( \begin{array}{cc} -1 & 0 \\ 1 & 0 \end{array} \right), \quad L_\beta = \left( \begin{array}{cc} 0 & 1 \\ 0 & -1 \end{array} \right).
\]

In this way, the general rate matrix for a continuous-time Markov chain on two states can be expressed as the linear combination\(^2\)

\[
Q = \alpha L_\alpha + \beta L_\beta = \left( \begin{array}{cc} -\alpha & \beta \\ \alpha & -\beta \end{array} \right).
\]

The associated substitution matrix $M(t)$ with matrix elements $[M(t)]_{ij}$, representing the probability of a substitution $j \to i$ after time $t$, is then given by the exponential map

\[
M(t) = \exp[Qt] := \sum_{m=0}^{\infty} \frac{Q^m t^m}{m!}.
\]

By noting that the generators satisfy the relations

\[
L_\alpha^2 = L_\beta L_\alpha = -L_\alpha, \quad L_\beta^2 = L_\alpha L_\beta = -L_\beta, \quad (1)
\]

it is easy to show that

\[
(\alpha L_\alpha + \beta L_\beta)^m = (-1)^m (\alpha + \beta)^m (\alpha L_\alpha + \beta L_\beta).
\]

Thus,

\(^1\)We use “Dirac notation”, where a vector is represented by a “ket” $|\cdot\rangle$, as this notation is particularly elegant when it comes to more general phylogenetic character patterns.

\(^2\)The reader should note that our rate matrices have zero column-sum rather than row-sum, as is sometimes more usual. The row-sum convention can be recovered from our presentation by taking the transpose of all matrices and replacing left multiplication with right.
\[ M(t) = \exp[Qt] = \sum_{m=0}^{\infty} \frac{Q^m t^m}{m!} = I - \frac{1}{\alpha + \beta} (e^{-(\alpha + \beta)t} - 1)(\alpha L_\alpha + \beta L_\beta) \]

\[ = I - \frac{1}{\alpha + \beta} (e^{-(\alpha + \beta)t} - 1)Q. \] (2)

**Remark 1** As \( M(t) \) is invariant under the re-parameterization\(^3\)

\[ t \to t' = \lambda t, \quad \alpha \to \alpha' = \lambda^{-1} \alpha, \quad \beta \to \beta' = \lambda^{-1} \beta, \]

we see that we can “scale out” \( t \) by choosing \( \lambda = t^{-1} \). As, in a practical context, \( \alpha, \beta, \) and even \( t \) are unknown parameters that must be inferred from observed data using some statistical estimation procedure; we see that we can take

\[ M(\alpha, \beta) = e^Q = e^{(\alpha L_\alpha + \beta L_\beta)}, \]

as completely equivalent to (2). If we think of \( \{M(\alpha, \beta) \mid \alpha, \beta \in \mathbb{R}\} \) as a two-dimensional manifold (in the sense of a Lie group: Procesi 2007), then we see that the Markov generators are none other than the basis vectors of the tangent space at the identity:

\[ L_\alpha \equiv \left. \frac{\partial}{\partial \alpha} M(\alpha, \beta) \right|_{\alpha = \beta = 0}, \quad L_\beta \equiv \left. \frac{\partial}{\partial \beta} M(\alpha, \beta) \right|_{\alpha = \beta = 0}, \]

with algebraic closure under the “Lie bracket” \([L_\alpha, L_\beta] := L_\alpha L_\beta - L_\beta L_\alpha = L_\alpha - L_\beta\) ensuring “closure” of the corresponding Markov model (as is discussed in Jarvis and Sumner 2010). This connection between continuous-time Markov chains and Lie groups is an important one and seems to have been first noted by Johnson (1985). This point of view is needed in order to extend the results of the present article to the case of character state spaces of arbitrary size. Having given this perspective into the meaning of the Markov generators, we will nevertheless take the more usual representation (2) of substitution matrices in all that follows below.

In Sumner and Jarvis (2005), it was shown that the Markov models of phylogenetics in standard use can be represented in an abstract setting using the tensor product space \( V \otimes V \otimes \cdots \otimes V \), where \( \dim(V) = k \) is the number of character states and the number of copies of \( V \) is equal to the number of taxa under consideration. In these models, it is usual to impose conditional independence across the branches of the tree and this can be formalized using a linear operator \( \delta : V \to V \otimes V \) to generate speciation events. This is referred to as the “splitting operator” and is defined, using our chosen basis, as

\[ \delta \cdot |i\rangle = |i\rangle \otimes |i\rangle, \]

\(^3\)For further discussion of local time-re-parameterization in phylogenetics, see Jarvis et al. (2005) or, in the context of a changing rate of mutation, see Penny (2005).
which, expressed at the level of the probability distributions, corresponds exactly to the duplication of a molecular unit. That is, if we select state $i$ from the initial sequence with probability $p_i$, then immediately after duplication—and assuming the sequences remain aligned—the probability of observing the pattern $ij$ at a given site is $p_i \delta_{ij}$. By defining the vector $p := \sum_i p_i |i\rangle$ and noting that the tensors $|ij\rangle := |i\rangle \otimes |j\rangle$ form a basis for $V \otimes V$, the splitting operator achieves this notion of speciation in the abstract setting:

$$
\delta \cdot p = \delta \cdot \left( \sum_i p_i |i\rangle \right) = \sum_i p_i \delta \cdot |i\rangle = \sum_i p_i |ii\rangle = \sum_{i,j} p_i \delta_{ij} |ij\rangle,
$$

where $P := \delta \cdot p \in V \otimes V$ has components $p_i \delta_{ij}$ and is referred to as a “phylogenetic tensor”. Subsequently, returning to the binary $k = 2$ case, given two rate matrices $Q_1$ and $Q_2$, and two edge weights $\tau_1$ and $\tau_2$, the phylogenetic tensor evolves to

$$
P' = e^{Q_1 \tau_1} \otimes e^{Q_2 \tau_2} \cdot P,
$$

which, up to first-order terms in the edge weights, is

$$
P' = \left[ (I + \alpha_1 \tau_1 L_\alpha + \beta_1 \tau_1 L_\beta + \cdots) \otimes (I + \alpha_2 \tau_2 L_\alpha + \beta_2 \tau_2 L_\beta + \cdots) \right] \cdot P
$$

$$
= \left[ I \otimes I + \alpha_1 \tau_1 L_\alpha \otimes I + \alpha_2 \tau_2 I \otimes L_\alpha + \beta_1 \tau_1 L_\beta \otimes I + \beta_2 \tau_2 I \otimes L_\beta + \cdots \right] \cdot P,
$$

where $\alpha_i$ and $\beta_i$ are the rate parameters appearing in $Q_i$. In this way, the splitting operator can be thought of as the generator of the branching pattern of the phylogenetic tree, while $L_\alpha$ and $L_\beta$ are the generators of the Markov process. (For more details of this formalism, see Bashford et al. 2004; Sumner and Jarvis 2005, and Sumner et al. 2008, and for an even more general setting see Jarvis et al. 2005.) Presently, we are concerned with the algebra resulting from application of these two types of generators.

### 3 Some Helpful Lemmas

The action of the Markov generators on the basis vectors is

$$
L_\alpha |0\rangle = |1\rangle - |0\rangle, \quad L_\beta |0\rangle = 0,
$$

$$
L_\alpha |1\rangle = 0; \quad L_\beta |1\rangle = |0\rangle - |1\rangle.
$$

By comparing the relations

$$
\delta \cdot L_\alpha |0\rangle = |11\rangle - |00\rangle, \quad \delta \cdot L_\alpha |1\rangle = 0;
$$

and

$$
L_\alpha \otimes L_\alpha |00\rangle = |11\rangle - |01\rangle - |10\rangle + |00\rangle,
$$

$$
L_\alpha \otimes L_\alpha |11\rangle = 0,
$$

$$
L_\alpha \otimes I |00\rangle = |10\rangle - |00\rangle,
$$

$$
L_\alpha \otimes I |00\rangle = |10\rangle - |00\rangle.
$$
\[ L_\alpha \otimes I |11\rangle = 0, \]
\[ I \otimes L_\alpha |00\rangle = |01\rangle - |00\rangle, \]
\[ I \otimes L_\alpha |11\rangle = 0; \]

with similar formulae for \( L_\beta \), we find that the Markov generators can be “pushed through” past the splitting operator:

**Lemma 3.1** As operators from \( V \) to \( V \otimes V \), we have

\[
\delta \cdot L_\alpha = (L_\alpha \otimes L_\alpha + L_\alpha \otimes I + I \otimes L_\alpha) \cdot \delta, \\
\delta \cdot L_\beta = (L_\beta \otimes L_\beta + L_\beta \otimes I + I \otimes L_\beta) \cdot \delta. 
\]

In the terminology of group actions (Procesi 2007), this lemma tells us the rule for how the two operators “intertwine”. It is exactly this relation that we will exploit to show how to generalize from a Markov model on a tree to a model on a general cluster system.

Given a linear operator \( X \) on \( V \), we define a linear operator \( X^{(i)} \) on \( V \otimes V \otimes \cdots \otimes V \) as the tensor product

\[
X^{(i)} := I \otimes I \otimes \cdots \otimes X \otimes I \otimes \cdots \otimes I, 
\]

where \( X \) appears in the \( i \)th slot of the tensor product. Further, for a subset \( A \subseteq [n] := \{1, 2, \ldots, n\} \), we define

\[
X^{(A)} := \prod_{i \in A} X^{(i)}. 
\]

For example, if we take \( n = 5 \), we have

\[
X^{(\{2, 5\})} = (I \otimes X \otimes I \otimes I \otimes I) \cdot (I \otimes I \otimes I \otimes I \otimes X) = I \otimes X \otimes I \otimes I \otimes X. 
\]

Presently, we will show how the interaction of \( \delta \) with \( L_\alpha \) naturally produces terms such as \( L_\alpha^{(A)} \) (and similar for \( L_\beta \)).

**Lemma 3.2** As linear operators from \( V \) to \( V \otimes V \otimes V \),

\[
I \otimes \delta \cdot \delta = \delta \otimes I \cdot \delta. 
\]

**Proof** We have

\[
I \otimes \delta \cdot \delta \cdot |i\rangle = I \otimes \delta \cdot (|i\rangle \otimes |i\rangle) = |i\rangle \otimes (|i\rangle \otimes |i\rangle) = |iii\rangle. 
\]

Similarly,

\[
\delta \otimes I \cdot \delta \cdot |i\rangle = \delta \otimes I \cdot (|i\rangle \otimes |i\rangle) = (|i\rangle \otimes |i\rangle) \otimes |i\rangle = |iii\rangle. 
\]

\( \square \)
Using this lemma, we can recursively define
\[
\delta^{i+1} := \delta \otimes I \otimes I \otimes \cdots \otimes I \cdot \delta^i
\]
\[
\equiv I \otimes \delta \otimes I \otimes \cdots \otimes I \cdot \delta^i \equiv \cdots \equiv I \otimes I \otimes \cdots I \otimes \delta \cdot \delta^i,
\]
with \(\delta^1 := \delta\). The action of the operator \(\delta^{n-1}\) taking \(V\) to \(V \otimes V \otimes \cdots \otimes V\) generates exactly the “\(n\)-taxon process” as defined in Bryant (2009) which, in turn, is completely equivalent to the formalism given in Bashford et al. (2004).

If we note that \(\delta \cdot I = I \otimes I \cdot \delta\) and consider Lemma 3.1, we see that, for \(x = \alpha\) or \(\beta\), we have
\[
\delta^2 \cdot L_x := \delta \otimes I \cdot \delta \cdot L_x
\]
\[
= \delta \otimes I \cdot (L_x \otimes L_x + I \otimes L_x + L_x \otimes I) \cdot \delta
\]
\[
= \left[(L_x \otimes L_x \otimes L_x + L_x \otimes I \otimes L_x + I \otimes L_x \otimes L_x) + I \otimes I \otimes L_x + (L_x \otimes L_x \otimes I + L_x \otimes I \otimes I + I \otimes L_x \otimes I)\right] \cdot \delta^2
\]
\[
= \left( \sum_{A \subseteq \{1,2,3\}, A \neq \emptyset} L_x^{(A)} \right) \cdot \delta^2.
\]
Generalizing this result, we have

**Lemma 3.3**

\[
\delta^{n-1} \cdot L_x = \left( \sum_{A \subseteq [n], A \neq \emptyset} L_x^{(A)} \right) \cdot \delta^{n-1}.
\]

**Proof** The proof is by induction. We have shown that the result is true for \(n = 3\). Assuming that it is true for some \(n > 3\), we have
\[
\delta^n \cdot L_x = (\delta \otimes I \otimes \cdots \otimes I) \cdot \left( \sum_{A \subseteq [n], A \neq \emptyset} L_x^{(A)} \right) \cdot \delta^{n-1}
\]
\[
= (\delta \otimes I \otimes \cdots \otimes I) \cdot \left( \sum_{A \subseteq [n-1], A \neq \emptyset} I \otimes L_x^{(A)} + \sum_{A \subseteq [n-1]} L_x \otimes L_x^{(A)} \right) \cdot \delta^{n-1}
\]
\[
= \left( \sum_{A \subseteq [n-1], A \neq \emptyset} I \otimes I \otimes L_x^{(A)} \right)
\]
\[
+ \sum_{A \subseteq [n-1]} (L_x \otimes L_x \otimes L_x^{(A)} + L_x \otimes I \otimes L_x^{(A)} + I \otimes L_x \otimes L_x^{(A)}) \right) \cdot \delta^n
\]
\[
= \left( \sum_{A \subseteq [n+1], A \neq \emptyset} L_x^{(A)} \right) \cdot \delta^n.
\]
\(\square\)
Recall that for \( n \geq 1 \), we have
\[
(\alpha L_\alpha + \beta L_\beta)^m = (-1)^{m-1}(\alpha + \beta)^{m-1}(\alpha L_\alpha + \beta L_\beta).
\] (3)

If we define
\[
\mathcal{L}_x^{[n]} := \sum_{A \subseteq [n], A \neq \emptyset} L_x^{(A)},
\] (4)
then Lemma 3.3 implies that we have the intertwining
\[
\delta^{n-1} \cdot L_x = \mathcal{L}_x^{[n]} \cdot \delta^{n-1}.
\]

We also note that we have the recursion
\[
\mathcal{L}_x^{[n]} = L_x \otimes \mathcal{L}_x^{[n-1]} + I \otimes \mathcal{L}_x^{[n-1]} + L_x \otimes I \otimes I \otimes \cdots \otimes I,
\]
and (after a little effort) it follows by induction on \( n \) that
\[
\left( \mathcal{L}_\alpha^{[n]} \right)^2 = \mathcal{L}_\beta^{[n]} \mathcal{L}_\alpha^{[n]} = -\mathcal{L}_\alpha^{[n]}, \quad \left( \mathcal{L}_\beta^{[n]} \right)^2 = \mathcal{L}_\alpha^{[n]} \mathcal{L}_\beta^{[n]} = -\mathcal{L}_\beta^{[n]}.
\]

Inspection reveals that, for all \( n \), \( \mathcal{L}_\alpha^{[n]} \) and \( \mathcal{L}_\beta^{[n]} \) satisfy exactly the same algebraic relations as \( L_\alpha \) and \( L_\beta \) that were given in (1). It follows immediately from (3) that for \( n \geq 1 \) we have
\[
(\alpha \mathcal{L}_\alpha^{[n]} + \beta \mathcal{L}_\beta^{[n]})^m = (-1)^{m-1}(\alpha + \beta)^{m-1}(\alpha \mathcal{L}_\alpha^{[n]} + \beta \mathcal{L}_\beta^{[n]}),
\]
and
\[
\delta^{n-1} \cdot (\alpha L_\alpha + \beta L_\beta)^m = (-1)^{m-1}(\alpha + \beta)^{m-1} \delta^{n-1} \cdot (\alpha L_\alpha + \beta L_\beta)
\]
\[
= (-1)^{m-1}(\alpha + \beta)^{m-1}(\alpha \mathcal{L}_\alpha^{[n]} + \beta \mathcal{L}_\beta^{[n]}) \cdot \delta^{n-1}
\]
\[
= (\alpha \mathcal{L}_\alpha^{[n]} + \beta \mathcal{L}_\beta^{[n]})^m \cdot \delta^{n-1}.
\]

Putting this together, we see that

**Lemma 3.4**
\[
\delta^{n-1} \cdot \exp[\alpha L_\alpha + \beta L_\beta] = \exp[\alpha \mathcal{L}_\alpha^{[n]} + \beta \mathcal{L}_\beta^{[n]}] \cdot \delta^{n-1}.
\]

In order to put all of the above to work, we require one final lemma regarding tensor products and the exponential map.

**Lemma 3.5** *Given any two linear operators \( X \) and \( Y \), we have*
\[
e^X \otimes e^Y = e^{X \otimes I + I \otimes Y}.
\]
Proof Consider
\[ e^X \otimes I = \left( I + X + \frac{1}{2}X^2 + \cdots \right) \otimes I \]
\[ = I \otimes I + X \otimes I + \frac{1}{2}(X \otimes I)^2 + \cdots = e^{X \otimes I}, \]
which implies that
\[ e^X \otimes e^Y = e^X \otimes I \cdot I \otimes e^Y = e^{X \otimes I} \cdot e^{I \otimes Y} = e^{X \otimes I + I \otimes Y}, \]
where the last identity follows because \( X \otimes I \) and \( I \otimes Y \) commute. □

4 Alternative Presentation of Markov Models on Trees

Consider the tree presented in Fig. 2. Suppose we are given a root distribution \( |\pi\rangle := \sum_i \pi_i |i\rangle \), a rate matrix \( Q = \alpha L_\alpha + \beta L_\beta \), and edge weights \( \tau_1, \tau_2, \tau_3, \tau_{34} \) and \( \tau_{234} \). The phylogenetic tensor corresponding to this tree can be generated as
\[ P = e^{Q_{\tau_1}} \otimes e^{Q_{\tau_2}} \otimes e^{Q_{\tau_3}} \otimes e^{Q_{\tau_{34}}} \cdot I \otimes I \otimes I \otimes \delta \cdot I \otimes I \otimes e^{Q_{\tau_{234}}} \cdot \delta \cdot |\pi\rangle. \]

If we set \( P = \sum_{i,j,k,l} p_{ijkl} |ijkl\rangle \) and interpret \( p_{ijkl} \) as the probability of observing the pattern \( ijk \) at the leaves of the tree, we see that this tensor is equivalent to specifying a joint distribution in the normal way (see Sumner and Jarvis 2005 for technical details of how this construction is equivalent to that discussed in Semple and Steel (2003, Chap. 8)).

By setting \( Q = \alpha L_\alpha + \beta L_\beta \) and applying Lemma 3.4, we find that
\[ I \otimes I \otimes \delta \cdot I \otimes I \otimes e^{Q_{\tau_{34}}} = I \otimes I \otimes (\delta \cdot e^{Q_{\tau_{34}}}) \]
\[ = I \otimes I \otimes e^{(\alpha L_\alpha + \beta L_\beta) \tau_{34}} \cdot I \otimes I \otimes \delta. \]

Now, we note that
\[ I \otimes I \otimes \delta \cdot I \otimes \delta = \left[ I \otimes (I \otimes \delta) \right] \cdot I \otimes \delta = I \otimes (I \otimes \delta \cdot \delta) = I \otimes \delta^2, \]

Fig. 2 A rooted tree on four leaves
so again applying Lemma 3.4 we find that
\[ I \otimes \delta^2 \cdot I \otimes e^{Q_{234}} = I \otimes e^{(\alpha L_3[3] + \beta L_3[3])} \tau_{234} \cdot I \otimes \delta^2. \]

We recall that
\[ I \otimes \delta^2 \cdot \delta = \delta^3. \]

Thus,
\[ P = e^{Q_{\tau_1}} \otimes e^{Q_{\tau_2}} \otimes e^{Q_{\tau_3}} \otimes e^{Q_{\tau_4}} \cdot I \otimes I \otimes e^{(\alpha L_2[2] + \beta L_2[2])} \tau_{234} \].

with \(|\delta^3 \pi\) := \delta^3 \cdot |\pi\) = \pi_0|0000\) + \pi_1|1111\). Finally, by applying Lemma 3.5 multiple times, we find that we can write
\[ P = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{34} R_{34}] \cdot \exp[\tau_{234} R_{234}] \cdot |\delta^3 \pi\),

where
\[ R_1 = \alpha (L_\alpha \otimes I \otimes I \otimes I) + \beta (L_\beta \otimes I \otimes I \otimes I), \]
\[ R_2 = \alpha (I \otimes L_\alpha \otimes I \otimes I) + \beta (I \otimes L_\beta \otimes I \otimes I), \]
\[ R_3 = \alpha (I \otimes I \otimes L_\alpha \otimes I) + \beta (I \otimes I \otimes L_\beta \otimes I), \]
\[ R_4 = \alpha (I \otimes I \otimes I \otimes L_\alpha) + \beta (I \otimes I \otimes I \otimes L_\beta), \]
\[ R_{34} = \alpha (I \otimes I \otimes L_\alpha[3]) + \beta (I \otimes I \otimes L_\beta[3]), \]
\[ R_{234} = \alpha (I \otimes L_\alpha[3]) + \beta (I \otimes L_\beta[3]). \]

Now consider the rooted quartet tree of Fig. 3. The phylogenetic tensor of this tree is given by
\[ P' = e^{Q_{\tau_1}} \otimes e^{Q_{\tau_2}} \otimes e^{Q_{\tau_3}} \otimes e^{Q_{\tau_4}} \cdot \delta \otimes \delta \cdot e^{Q_{\tau_{12}}} \otimes e^{Q_{\tau_{34}}} \cdot \delta \cdot |\pi\).\]

By a similar argument to the one just given, it is possible to show that this tensor can be re-expressed as
\[ P' = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{12} R_{12} + \tau_{34} R_{34}] \cdot |\delta^3 \pi\),

Fig. 3 Alternative tree on four leaves
with
\[
R_{12} = \alpha (\mathcal{L}_\alpha^{[2]} \otimes I \otimes I) + \beta (\mathcal{L}_\beta^{[2]} \otimes I \otimes I).
\]

We can extend our definition (4) of \( \mathcal{L}_x^{[n]} \) to arbitrary subsets by taking
\[
\mathcal{L}_x^A := \sum_{B \subseteq A, B \neq \emptyset} L_x^{(B)},
\]
for all \( A \subseteq [n] \). Now label the elements of \( A \) as \( A = \{a_1, a_2, \ldots, a_{|A|}\} \) and consider a permutation \( \sigma \in \mathfrak{S}_n \) such that \( \sigma(a_i) = i \) (obviously such a permutation always exists). If we allow \( \sigma \) to act on \( V^{\otimes n} \) by permuting tensor factors, it is clear that
\[
\sigma(\mathcal{L}_x^A) = \mathcal{L}_x^{|A|} \otimes I^{([n]-|A|)}.
\]

From this, we can conclude that for fixed \( A \) the operators \( \mathcal{L}_\alpha^A \) and \( \mathcal{L}_\beta^A \) satisfy the same algebraic relations as \( L_\alpha \) and \( L_\beta \):
\[
(\mathcal{L}_\alpha^A)^2 = \mathcal{L}_\beta^A \mathcal{L}_\alpha^A = -\mathcal{L}_\alpha^A, \quad (\mathcal{L}_\beta^A)^2 = \mathcal{L}_\alpha^A \mathcal{L}_\beta^A = -\mathcal{L}_\beta^A,
\]
for all \( A \subseteq [n] \). Using this result, we can unify the expressions for the rate matrices above by defining
\[
R_A := \alpha \mathcal{L}_\alpha^A + \beta \mathcal{L}_\beta^A.
\]

Evidently, \( A_1 \cap A_2 = \emptyset \) implies that \([R_{A_1}, R_{A_2}] = 0\) and we see that there are several ways we can express our two phylogenetic tensors. For instance, the following presentations of the tree given in Fig. 2 are all equally valid:
\[
P = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{34} R_{34}] \cdot \exp[\tau_{234} R_{234}] \cdot |\delta^3 \pi|
\]
\[
= \exp[\tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_1 R_1 + \tau_{34} R_{34}] \cdot \exp[\tau_{234} R_{234}] \cdot |\delta^3 \pi|
\]
\[
= \exp[\tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{34} R_{34}] \cdot \exp[\tau_1 R_1 + \tau_{234} R_{234}] \cdot |\delta^3 \pi|
\]
\[
= \exp[\tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{34} R_{34}] \cdot \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_{234} R_{234}] \cdot |\delta^3 \pi|.
\]

The content of our main theorem is that there is a canonical choice of presentation of a phylogenetic tensor for arbitrary trees. Consider the particular presentations of the quartet trees of Figs. 2 and 3, respectively:
\[
P = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{34} R_{34}] \cdot \exp[\tau_{234} R_{234}] \cdot |\delta^3 \pi|,
\]
\[
P' = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3 + \tau_4 R_4] \cdot \exp[\tau_{12} R_{12} + \tau_{34} R_{34}] \cdot |\delta^3 \pi|.
\]

**Theorem 4.1** Consider a rooted tree \( T \) with \( n \) leaves represented as the cluster system \( T = \{A_1, A_2, \ldots, A_{2n+2}\} \), where \( A_i \subset [n] \). Given a root distribution \( \pi \), any rate parameters \( \alpha \) and \( \beta \) and weights \( \{\tau_{A_1}, \tau_{A_2}, \ldots, \tau_{A_{2n+2}}\} \), a phylogenetic tensor (or joint distribution) \( P \) at the leaves of this cluster system can be expressed as
\[
P = \exp[\mathcal{X}_1] \cdot \exp[\mathcal{X}_2] \cdot \cdots \cdot \exp[\mathcal{X}_{n-1}] \cdot |\delta^{n-1} \pi|,
\]
with
\[ \delta^{n-1} \pi := \delta^{n-1} \cdot |\pi\rangle = \pi_0|00\ldots0\rangle + \pi_1|11\ldots1\rangle, \]

and
\[ X_i = \sum_{A \in \mathcal{T}, |A| = i} \tau_A \mathcal{R}_A. \]

**Proof** The proof is by induction. Clearly a phylogenetic tensor on two leaves can be placed into the required form:
\[ P = \exp[\tau_1 \mathcal{R}_1 + \tau_2 \mathcal{R}_2] \cdot |\delta\pi\rangle. \]

Assume that for \( k > 2 \), any phylogenetic tensor on \( k \) leaves \( P^{(k)} \) can be expressed in the required form:
\[ P^{(k)} = \exp[\tau_1 \mathcal{R}_1 + \tau_2 \mathcal{R}_2 + \cdots + \tau_n \mathcal{R}_n] \cdot \exp\left[ \sum_{A, |A| = 2} \tau_A \mathcal{R}_A \right] \cdots \cdot \exp\left[ \sum_{A, |A| = n-1} \tau_A \mathcal{R}_A \right] \cdot |\delta^{k-1}\pi\rangle. \]

Without loss of generality, we generate a phylogenetic tensor on \( k + 1 \) leaves, \( P^{(k+1)} \), by branching \( P^{(k)} \) at the leaf \( k \). This is expressed algebraically by
\[ P^{(k)} \rightarrow P^{(k+1)} = (I \otimes I \otimes \cdots \otimes I \otimes \delta) \cdot P^{(k)}. \]

For an arbitrary subset \( A \subset [k] \), we have
\[ I \otimes I \otimes \cdots \otimes I \otimes \delta \cdot \mathcal{R}_A = \mathcal{R}_{A'}. \]

where
\[ A' = \begin{cases} A \cup \{k + 1\}, & \text{if } k \in A, \\ A, & \text{otherwise.} \end{cases} \]

Pushing right through, we have
\[ P^{(k+1)} = \exp\left[ \sum_{A, |A| = 1} \tau_A \mathcal{R}_{A'} \right] \cdot \exp\left[ \sum_{A, |A| = 2} \tau_A \mathcal{R}_{A'} \right] \cdots \cdot \exp\left[ \sum_{A, |A| = n-1} \tau_A \mathcal{R}_{A'} \right] \cdot |\delta^k\pi\rangle. \]

Now consider, for a given \( 1 \leq i < k - 1 \), the term
\[ \exp\left[ \sum_{A, |A| = i} \tau_A \mathcal{R}_{A'} \right]. \]
As \( k \in A \) for only one cluster \( A \) in this summation, there is exactly one \( A' \) of cardinality \( i+1 \) in it. We label this subset as \( B_{i+1} \) and its corresponding edge weight as \( \tau_{B_i} \), so that \( B_{i+1} = B_i \cup \{k+1\} \). Additionally, \( B_{i+1} \) is disjoint from all of the other subsets, so we may write

\[
\exp \left[ \sum_{A, |A| = i, A \neq \emptyset} \tau_A R_A \right] = \exp \left[ \sum_{A, |A| = i+1, A \neq \emptyset, A \neq B_i} \tau_A R_A \right] \exp[\tau_{B_i} R_{B_i+1}].
\]

Exactly the same argument is valid for the \( i+1 \) term:

\[
\exp \left[ \sum_{A, |A| = i+1, A \neq \emptyset} \tau_A R_A' \right] = \exp \left[ \sum_{A, |A| = i+1, A \neq \emptyset, A \neq B_{i+2}} \tau_A R_A \right] \exp[\tau_{B_{i+1}} R_{B_{i+2}}].
\]

Thus, we can express the product of these two terms as

\[
\exp \left[ \sum_{A, |A| = i} \tau_A R_A' \right] \cdot \exp \left[ \sum_{A, |A| = i+1} \tau_A R_A' \right] = \exp \left[ \sum_{A, |A| = i, A \neq \emptyset, A \neq B} \tau_A R_A \right] \exp[\tau_{B_i} R_{B_{i+1}} + \sum_{A, |A| = i+1, A \neq \emptyset, A \neq B_{i+2}} \tau_A R_A ]
\]

\[
\cdot \exp[\tau_{B_{i+1}} R_{B_{i+2}}].
\]

Continuing in this way we can place \( P^{(k+1)} \) in the required form and the theorem follows by induction on \( k \). 

From a biological perspective, it is apparent that the form given in Theorem 4.1 that utilizes a cardinality ordering is somewhat mysterious as there is not immediately a biological motivation for branching events to occur in an order that respects this ordering. However, a little thought using commutativity (or otherwise) of the various operators shows that it is not so much the cardinality that matters, but it is that the operators that arise independently across branches of the tree are necessarily commutative and, conversely, those that do not commute necessarily have non-zero intersection, and hence are not independent. Noting that there is some freedom in the final expression, we see that the cardinality ordering is simply a nice way of unifying the description for arbitrary trees. In fact other orderings are possible such as the topological depth of node in the tree, as measured from root, or by the maximum topological depth of node in the tree, as measured from the leaves.

It is clear that if we take \( P \) as in Theorem 4.1 in the case that the cluster system \( T \) is compatible, we have a probability distribution on a tree identical to the standard presentation usually given in phylogenetics (Semple and Steel 2003, Chap. 8). However, the construction we have given naturally generalizes to a model on general cluster systems with trees occurring as sub-models where the weights associated to
clusters incompatible with the given tree are set to zero. There is also an obvious
generalization to the case where each cluster has a unique rate matrix—simply give
additional cluster labels to the rate parameters: $\alpha \to \alpha_A$ and $\beta \to \beta_A$.

4.1 Extension to the $n$-State General Markov Model

For the $n$-state general Markov model, we can write the generic rate-matrix as

$$Q = \sum_{1 \leq i \neq j \leq n} \alpha_{ij} L_{ij},$$

where, for each $i \neq j$, $L_{ij}$ is the $n \times n$ matrix with 1 in the $ij$ entry, $-1$ in the $jj$
entry and zeros elsewhere. The matrices $L_{ij}$ satisfy the algebraic relations

$$L_{ij} L_{kl} = \delta_{jk} (L_{il} - L_{jl}) - \delta_{jl} L_{il}.$$

Taking $V \cong \mathbb{C}^n$ with basis $\{|i\rangle\}_{1 \leq i \leq n}$, we generalize Lemma 3.1 for the $n$-state model:

**Lemma 4.2** As operators from $V$ to $V \otimes V$, we have

$$\delta \cdot L_{ij} = (L_{ij} \otimes L_{ij} + L_{ij} \otimes I + I \otimes L_{ij}) \cdot \delta.$$

**Proof** Recalling the structure of each $L_{ij}$, we see from direct computation that

$$L_{ij} |k\rangle = \delta_{jk} (|i\rangle - |k\rangle).$$

Thus, we have

$$\delta \cdot L_{ij} := \delta_{jk} (|i\rangle - |k\rangle) = \delta_{jk} (|ii\rangle - |kk\rangle).$$

On the other hand,

$$(L_{ij} \otimes L_{ij} + L_{ij} \otimes I + I \otimes L_{ij}) \cdot \delta \cdot |k\rangle$$

$$= (L_{ij} \otimes L_{ij} + L_{ij} \otimes I + I \otimes L_{ij}) |kk\rangle$$

$$= \delta_{jk} (|i\rangle - |k\rangle) \otimes (|i\rangle - |k\rangle) + |k\rangle \otimes (|i\rangle - |k\rangle) + (|i\rangle - |k\rangle) \otimes |k\rangle$$

$$= \delta_{jk} (|ii\rangle - |kk\rangle).$$

Armed with this result, it is straightforward to revisit our previous lemmas to show
that they remain valid for the $n$-state case. In this way, we see that we can generalize
Theorem 4.1 to the $n$-state general Markov model.

Theorem 4.1 represents the main result of this article. In the next section, we
specialize to the binary-symmetric model with $\alpha = \beta$ in order to make connections
with previous work.
5 Comparison with the Hadamard Transformation of the Binary-Symmetric Model

In this section, we consider the “binary-symmetric” case where $\alpha = \beta = 1$, so that we can write
\[
Q = (L_\alpha + L_\beta) = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} = -I + K,
\] (5)
where $K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is the permutation matrix taking $|0\rangle \leftrightarrow |1\rangle$.

**Lemma 5.1** If $K_\sigma$ is a permutation matrix representing the permutation $\sigma$, i.e. $K_\sigma |i\rangle = |\sigma(i)\rangle$, then
\[
\delta \cdot K_\sigma = K_\sigma \otimes K_\sigma \cdot \delta.
\]

**Proof** Clearly,
\[
\delta \cdot K_\sigma |i\rangle = \delta \cdot |\sigma(i)\rangle = |\sigma(i)\rangle \otimes |\sigma(i)\rangle = K_\sigma \otimes K_\sigma \cdot |i\rangle \otimes |i\rangle = K_\sigma \otimes K_\sigma \cdot \delta |i\rangle.
\]

Using Lemma 5.1, it is shown in Bashford et al. (2004) that for any tree $T$ represented as a cluster system we have
\[
P = \exp\left[ \sum_{x \in T} w_x (K^{(x)} - I \otimes I \otimes \cdots \otimes I) \right] |\delta^{n-1} \pi\rangle
\]
\[
= e^{-\lambda} \exp\left[ \sum_{x \in T} w_x K^{(x)} \right] |\delta^{n-1} \pi\rangle,
\]
where $\{w_x\}_{x \in T}$ is any set of edge weights on $T$ and $\lambda = \sum_{x \in T} w_x$. We will refer to this constructing of a phylogenetic tensor as the “$K$-representation”.

On the other hand, we have shown in Theorem 4.1 that for $\alpha = \beta = 1$ we can write
\[
P = \exp[\mathcal{X}_1] \cdot \exp[\mathcal{X}_2] \cdots \cdot \exp[\mathcal{X}_{n-1}] |\delta^{n-1} \pi\rangle,
\]
with
\[
\mathcal{X}_i = \sum_{A, |A| = i} \tau_A \mathcal{R}_A,
\]
and
\[
\mathcal{R}_A = \sum_{B \subseteq A, B \neq \emptyset} (\mathcal{L}_\alpha^B + \mathcal{L}_\beta^B).
\]
We will refer to this construction of a phylogenetic tensor as the “$\mathcal{L}$-representation”. We will show that for a tree these two representations are exactly equal, but for arbitrary cluster systems this is not the case.
We will find it convenient to label the tensor $|i_1i_2\ldots i_n\rangle$ by the subset $A \subset [n]$ defined by setting $j \in A$ if and only if $i_j = 1$. For example, if $n = 6$, we have

$$|\emptyset\rangle = |000000\rangle,$$
$$|[6]\rangle = |\{1, 2, 3, 4, 5, 6\}\rangle = |111111\rangle,$$
$$|[2, 3, 4]\rangle = |011100\rangle,$$
$$|[5]\rangle = |000010\rangle.$$

**Lemma 5.2** If

(a) $A \subseteq B$, it follows that $\mathcal{R}_A|B\rangle = |B - A\rangle - |B\rangle$,

(b) $A \cap B = \emptyset$, it follows that $\mathcal{R}_A|B\rangle = |B \cup A\rangle - |B\rangle$.

*In either case (a) or (b), we have

$$\mathcal{R}_A|B\rangle = (K^A - I \otimes I \otimes \cdots \otimes I)|B\rangle.$$

**Proof** For all $A \subseteq B$, we have

$$K^A|B\rangle = |B - A\rangle.$$

On the other hand we know that

$$\mathcal{R}_{[n]}|\emptyset\rangle = |[n]\rangle - |\emptyset\rangle,$$
$$\mathcal{R}_{[n]}|[n]\rangle = |\emptyset\rangle - |[n]\rangle.$$

For $A \subseteq B \subseteq [n]$, it is always possible to permute tensor factors to write

$$|B\rangle \cong |[n_A]\rangle \otimes |B'\rangle,$$

where $n_A = |A|$ and $B' \subset [n] - A$ with $B' := B - A$. Thus,

$$\mathcal{R}_A|B\rangle \cong (\mathcal{R}_{[n_A]}|[n_A]\rangle) \otimes |B'\rangle$$
$$= (|\emptyset\rangle - |[n_A]\rangle) \otimes |B'\rangle$$
$$\cong |B - A\rangle - |B - A \cup A\rangle$$
$$= |B - A\rangle - |B\rangle,$$

which proves the lemma for $A \subseteq B$. The $A \cap B = \emptyset$ case follows from a similar argument. \qed

For two subsets $A, B$ taken from a compatible cluster system with $|A| = |B|$ it is the case that $A \cap B = \emptyset$, which in turn implies that $[\mathcal{R}_A, \mathcal{R}_B] = 0$. Thus we can make the replacement

$$\exp[\lambda_i] := \exp\left[\sum_{A, |A| = i} \tau_A \mathcal{R}_A\right] = \prod_{A, |A| = i} \exp[\tau_A \mathcal{R}_A],$$
where by commutativity the product can be ordered in any way we please. Thus, in the $L$-representation, we see that for a tree we can write

$$P = \exp[\tau A_3 R A_3] \exp[\tau A_2 R A_2] \exp[\tau A_1 R A_1] \delta^{n-1} \pi,$$

with, due to the fact we are dealing with a tree, either $A_i \cap A_{i+1} = \emptyset$ or $A_{i+1} \subset A_i$. Noting that $\delta^{n-1} \pi = \pi_0 |\emptyset\rangle + \pi_1 |[n]\rangle$ and repeated application of Lemma 5.2 then gives the following theorem.

**Theorem 5.3** For compatible cluster systems, the “$L$-representation” and the “$K$-representation” give rise to identical phylogenetic tensors.

For arbitrary cluster systems, however, this is not true, as the example in the next section shows.

### 6 Phylogenetic Networks and Epochs

Consider the three-taxon phylogenetic tree given in Fig. 4(a). As was proved above in Theorem 5.3, if we take the binary symmetric model we get an identical probability distribution if we use either the $L$-representation $P_L = \exp[\tau_1 R_1 + \tau_2 R_2 + \tau_3 R_3] \exp[\tau_{12} R_{12}] \delta^{n-1} \pi$ or the $K$-representation $P_K = e^{-\lambda} \exp[\tau_1 K^{(1)} + \tau_2 K^{(2)} + \tau_3 K^{(3)} + \tau_{12} K^{(12)}] \delta^{2\pi}$, with $\lambda = \tau_1 + \tau_2 + \tau_3 + \tau_{12}$.

We would like to introduce the additional parameter $\tau_{23}$ associated with the incompatible cluster $\{2, 3\}$ to these probability distributions. We will do this in a way which is consistent with the design given in Fig. 4(b), where the evolutionary history in broken up into three epochs: E1. divergence of taxa 3 away from 1 and 2, E2. convergent evolution of taxa 2 and 3, with independent divergence of taxa 3, and E3. independent divergence of all taxa.

To model this situation, we must introduce the additional edge to each representation. To make the presentation as simple as possible, we impose a molecular clock on the model such that $\tau_2 = \tau_1$ and $\tau_3 = \tau_{12} + \tau_1$, and we introduce a scaling parameter $\theta \in [0, 1]$ to control the length of the second epoch as a proportion of the third epoch by setting $\tau_{23} = \tau_1 \theta$. As all operators in the $K$-representation commute, the only choice available in this case is to take...
Fig. 5 Substitution rates for states $|xyz\rangle$ under (a) $R_{23}$, and (b) $K^{(23)} - I \otimes I \otimes I$

\[P'_k = e^{-\lambda} \exp \left[ \tau_1 K^{(1)} + \tau_1 (1-\theta) K^{(2)} + (\tau_{12} + \tau_1 (1-\theta)) K^{(3)} \right.\]
\[\left. + \tau_{12} K^{(12)} + \tau_1 \theta K^{(23)} \right]\delta^2 \pi.\]

This is exactly consistent with the generalizations given in Bryant (2005, 2009).

For the $L$-representation, we do not have commutativity of the operators $R_2, R_3, R_{12}$ with the new operator $R_{23}$, thus we need to proceed more carefully as there is some choice in how the extra edge is introduced. Using the diagram of Fig. 4(b) and its three epochs as a guide, we take

\[P'_\ell = \exp \left[ \tau_1 (1-\theta)(R_1 + R_2 + R_3) \right] \cdot \exp[\tau_1 \theta (R_1 + R_{23})] \cdot \exp[\tau_{12}(R_3 + R_{12})]\delta^2 \pi.\]

For clarity of comparison, we write the $K$-representation in epoch form:

\[P'_k = e^{-\lambda} \exp \left[ \tau_1 (1-\theta) (K^{(1)} + K^{(2)} + K^{(3)}) \right] \cdot \exp[\tau_1 \theta (K^{(1)} + K^{(23)})] \cdot \exp[\tau_{12}(K^{(3)} + K^{(12)})]\delta^2 \pi.\]

Now, consider the state of the probability distribution at the beginning of epoch 2. As we are dealing with the binary symmetric model, it is clear that the probability of any state $|ijk\rangle$ is invariant to permutation of the states $0 \leftrightarrow 1$. Also, the structure of the tree up to the start of epoch 2 implies that the probability of any state of the form $|ijk\rangle$ where $i \neq j$ is of probability zero. Thus, we can assume at the start of epoch 2 that the distribution is of the form

\[P = (1 - q) \frac{1}{2}(|000\rangle + |111\rangle) + q \frac{1}{2}(|001\rangle + |110\rangle),\]

where, because we are considering a continuous-time model, $q \in [0, \frac{1}{2})$.

Considering the expansions

\[R_{23} = I \otimes (L_\alpha \otimes L_\alpha + L_\alpha \otimes I + I \otimes L_\alpha + L_\beta \otimes L_\beta + L_\beta \otimes I + I \otimes L_\beta),\]

and

\[K^{(23)} = I \otimes K \otimes K = R_{23} + I \otimes (L_\alpha \otimes L_\beta + L_\beta \otimes L_\alpha),\]

it follows that substitution rates between the four existing states in the two models are given by the two graphs in Fig. 5, where all substitution rates are equal. The crucial thing to note is that $R_{23}$ “corrects” patterns that are inconsistent with the cluster $\{2, 3\}$, whereas $K^{(23)}$ simply interchanges these two states.
It is now clear that there is quite a marked difference between the \( \mathcal{L} \)- and \( K \)-representations. The \( \mathcal{L} \)-representation introduces a natural notion of the “coming together” of taxa. In fact, it is easy to see directly from Fig. 5 that in the limit of extension of the edge \( \tau_{23} \) to infinity that the probability distribution will converge to

\[
\left( 1 - \frac{1}{2}q \right) \frac{1}{2}(|000\rangle + |111\rangle) + \frac{1}{2}q \frac{1}{2}(|011\rangle + |100\rangle),
\]

which is consistent with the probability distribution that would arise under a tree where taxa 1 has diverged from 2 and 3, but there has been zero divergence of taxa 2 and 3 themselves. This behaviour is of the course the reasoning behind the way we have chosen to draw our diagram Fig. 4(b).

The \( K \)-representation cannot achieve this type of convergence, with its limiting state being

\[
(1 - q) \frac{1}{4}(|000\rangle + |011\rangle + |111\rangle + |100\rangle) + q \frac{1}{4}(|001\rangle + |010\rangle + |110\rangle + |101\rangle).
\]

From this, we see that the \( K \)-representation is not capable of removing the patterns that are consistent with the tree 12|3. This is essentially the case because this representation is built up by expressing the rate matrix \( Q \) as a sum of permutation matrices (as in (5)) and applying Lemma 5.1. This observation generalizes immediately to the case of Kimura 3ST model, where the rate matrices can also be expressed as sums of permutation matrices (as is discussed in Bashford et al. 2004):

\[
Q = \begin{pmatrix}
* & \alpha & \beta & \gamma \\
\alpha & * & \gamma & \beta \\
\beta & \gamma & * & \alpha \\
\gamma & \beta & \alpha & *
\end{pmatrix} = -(\alpha + \beta + \gamma) I + \alpha K_\alpha + \beta K_\beta + \gamma K_\gamma,
\]

where

\[
K_\alpha = \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{pmatrix}, \quad K_\beta = \begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{pmatrix}, \quad K_\gamma = \begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{pmatrix}.
\]

Revisiting Lemma 3.1 further clarifies the distinction between the \( \mathcal{L} \)- and \( K \)-representations. The \( \mathcal{L} \)-representation maintains the separate identities of the “\( \alpha \)”- and “\( \beta \)”-rates as they intertwine with the splitting operator \( \delta \) and then we set \( \alpha = \beta \) to give the binary-symmetric case. By decomposing \( Q \) into permutation matrices and relying on Lemma 5.1, the \( K \)-representation compounds the two rates together irreversibly at the outset. We suspect that this is exactly the reason that a generalization to the general Markov model has previously proved elusive.

The ability of the \( \mathcal{L} \)-representation to model convergent evolution comes at a price, however. As is hinted at above, the resulting network model class is extremely general with the notion of “epochs” being essential to get a handle on what is possible. Essentially, one can now express models where the evolutionary history is split up into any number of epochs wherein the sequence evolution of arbitrary subsets of
taxa can be deemed to be diverging or converging. The case of trees corresponds exactly to the case where the start of each epoch is marked by the divergence of a lineage (identified as a subset of taxa) into a pair of non-intersecting subsets.

**Definition 6.1** A phylogenetic *epoch model* with $q$ epochs is given by an expression of the form

$$P = \exp \left( \sum_{A \subseteq [n], A \neq \emptyset} \theta_A^{(1)} R_A \right) \cdot \exp \left( \sum_{A \subseteq [n], A \neq \emptyset} \theta_A^{(2)} R_A \right) \cdots \times \exp \left( \sum_{A \subseteq [n], A \neq \emptyset} \theta_A^{(q)} R_A \right) \cdot |\delta^{n-1}\pi).$$

Theorem 4.1 shows that the general Markov model on a tree is a special example of an epoch model, as are the examples given in this section.

What should be highlighted in all of this is that there is a delicate interplay between the discrete structure of phylogenetic clusters and the commutativity (or otherwise) of the associated rate operators. Clearly, further research is required to elucidate the applicability in various biological circumstances of the models that our construction provides.

7 Conclusion

In this article, we have shown how to express the two-state continuous-time general Markov model on trees in such a way that allows extension to arbitrary cluster systems and even more general network models. By reviewing the lemmas in Sect. 2, it is clear that the results extend easily to the general Markov model with more character states. However, we defer confirmation of this observation to future work.

In Sect. 5, we showed that our discussion gives rise to phylogenetic models for the general Markov model that are identical to previous approaches only on trees, and in Sect. 6 we gave a simple example that shows how our approach allows for convergent evolution of previously divergent lineages (a structural property that was previously unobtainable).

Besides its theoretical interest, we expect that the ability to model convergent evolution in this way will have a significant application where it is known that particular datasets exhibit non-treelike behaviour due to population genetic properties such as incomplete lineage sorting and other effects that confound strictly treelike models (as discussed in Sect. 1). We suspect that exploration of the relation between the network models that arise in our discussion and simple models of population genetics is likely to yield significant additional insight.

Comparison of our network models to the distribution space generated by mixture models is also in need of investigation. For example, comparison of a mixture of the trees 12|3 and 1|23 to our network example given in Sect. 6 should yield significant theoretical insight into the biological meaning and plausibility of these differing model classes. Careful scientific thought is required to tease out what biological processes are explicitly (or implicitly) being modelled by either of these approaches.
Of course, these exciting possibilities must be tempered by analysis of the identifiability (or otherwise) of the models that arise by taking more general networks. Establishing identifiability is not only essential from a statistical inference point of view, but in this case, may lead to natural restrictions of the types of network that can be realistically used for phylogenetic inference.

Acknowledgements We would like to thank David Bryant for useful discussions. We would also like to thank the two anonymous reviewers for their careful reading of an early draft and resulting comments that have allowed for a greatly improved article. This research was support by ARC Discovery Project DP0877447 and Future Fellowship FT100100031.

References

Bandelt, H.-J. (1994). Phylogenetic networks. In Verhandlungen des Naturwissenschaftlichen Vereins Hamburg: Vol. 34.

Bandelt, H. J., & Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics and Evolution, 1, 242–252.

Bashford, J. D., Jarvis, P. D., Sumner, J. G., & Steel, M. A. (2004). U(1) × U(1) × U(1) symmetry of the Kimura 3ST model and phylogenetic branching processes. Journal of Physics. A, Mathematical and General, 37, L1–L9.

Bryant, D. (2005a). Extending tree models to split networks. In: L. Pachter & B. Sturmfels (Eds.), Algebraic statistics and computational biology (pp. 297–368). Cambridge: Cambridge University Press.

Bryant, D. (2005b). On the uniqueness of the selection criterion in Neighbor-Joining. Journal of Classification, 22, 3–15.

Bryant, D. (2009). Hadamard phylogenetic methods and the n-taxon process. Bulletin of Mathematical Biology, 71, 297–309.

Bryant, D., & Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution, 21, 255–265.

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

Holland, B., & Moulton, V. (2003). Consensus networks: a method for visualising incompatibilities in collections of trees. In: G. Benson & R. Page (Eds.), 3rd international workshop on algorithms in bioinformatics (WABI 2003) (pp. 165–176). Berlin: Springer.

Holland, B. R., Jermiin, L. S., & Moulton, V. (2006). Improved consensus network techniques for genome-scale phylogeny. Molecular Biology and Evolution, 23, 848–855.

Huson, D. H., Rupp, R., & Scornavacca, C. (2011). Phylogenetic networks: concepts, algorithms and applications. Cambridge: Cambridge University Press.

Jarvis, P. D., Bashford, J. D., & Sumner, J. G. (2005). Path integral formulation and Feynman rules for phylogenetic branching models. Journal of Physics. A, Mathematical and General, 38, 9621–9647.

Jarvis, P. D., & Sumner, J. G. (2010). Markov invariants for phylogenetic rate matrices derived from embedded submodels. arXiv:1008.1121, to appear.

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., & Larkum, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. BMC Systems Biology, 53, 638–643.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2006). Maximum likelihood of phylogenetic networks. Bioinformatics, 21, 2604–2611.

Johnson, J. E. (1985). Markov-type Lie groups in GL(n, R). Journal of Mathematical Physics, 26, 252–257.

Penny, D. (2005). Relativity for molecular clocks. Nature, 436, 183–184.

Procesi, C. (2007). Lie groups: an approach through invariants and representations. Berlin: Springer.

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

Strimmer, K., & Moulton, V. (2000). Likelihood analysis of phylogenetic networks using directed graphical models. Molecular Biology and Evolution, 17, 875–881.
Sumner, J. G., Charleston, M. A., Jermiin, L. S., & Jarvis, P. D. (2008). Markov invariants, plethysms, and phylogenetics. *Journal of Theoretical Biology, 253*, 601–615.

Sumner, J. G., & Jarvis, P. D. (2005). Entanglement invariants and phylogenetic branching. *Journal of Mathematical Biology, 51*, 18–36.

Von Haeseler, A., & Churchill, G. A. (1993). Network models for sequence evolution. *Journal of Molecular Evolution, 37*, 77–85.