Lie-Markov Models Derived from Finite Semigroups

Jeremy G. Sumner\(^1\) \cdot Michael D. Woodhams\(^1\)

Received: 12 September 2017 / Accepted: 8 June 2018 / Published online: 2 August 2018
© Society for Mathematical Biology 2018

Abstract

We present and explore a general method for deriving a Lie-Markov model from a finite semigroup. If the degree of the semigroup is \(k\), the resulting model is a continuous-time Markov chain on \(k\)-states and, as a consequence of the product rule in the semigroup, satisfies the property of multiplicative closure. This means that the product of any two probability substitution matrices taken from the model produces another substitution matrix also in the model. We show that our construction is a natural generalization of the concept of group-based models.

Keywords

Lie algebras \· Continuous-time Markov chains \· Group-based models \· Phylogenetics

1 Introduction

Recent work has defined and explored ‘Lie-Markov models’ in the context of phylogenetic modelling (Sumner et al. 2012a). These are the class of continuous-time Markov models which have the pleasing property of producing substitution matrices which are closed under matrix multiplication (Sumner 2017). We have argued this is an important consistency property for a phylogenetic model to possess (Sumner et al. 2012b) and have shown there is some evidence these models perform better than entrenched standard models on real data sets (Woodhams et al. 2015).

If one considers the definition of Lie-Markov models to be mathematically compelling in its own right, it is an interesting mathematical question to produce a complete enumeration of these models. Given the well-developed nature of the associated Lie group theory [see Hall (2015), Stillwell (2008) for excellent introductions], one might think it is simply a matter of looking in a standard reference to find a list of all Lie matrix groups and rejecting those which are not Markovian (in the appropriate sense). However, since the existing classifications of Lie matrix groups are stated up to iso-
morphism and, in particular, matrix similarity transformations, such an approach fails to be viable in practice.

As a practical alternative, the approach developed in Sumner et al. (2012a) rests upon the natural observation that the phylogenetic models that are used in practice have symmetries under nucleotide permutations. For example, the general time-reversible (GTR) model, which is shown to not be multiplicatively closed in Sumner et al. (2012a), nonetheless has complete symmetry under nucleotide permutations. In Sumner et al. (2012a), we presented a general method which produces all Lie-Markov models which have the symmetries of a given permutation group. This method was then applied to show there are five Lie-Markov models with complete symmetry. We applied the same method in Fernández-Sánchez et al. (2015) to show there are (approximately) 35 Lie-Markov models with symmetry that respects the partitioning of nucleotides in purines AG and pyrimidines CT. (The precise count depends on inclusion/exclusion of some special cases depending upon one’s preferences.) Currently, this is the most powerful method we know of for systematically deriving Lie-Markov models.

We recall that the literature contains two well-known classes of phylogenetic models that are special cases of Lie-Markov models: the so-called group-based (Semple and Steel 2003) and equivariant (Draisma and Kuttler 2009) models (see Sects. 3.4, 3.5 and, respectively). As they admit a Fourier analysis (Székely et al. 1993), group-based models have been extensively studied for both their attractive mathematical properties and biological applications [see, for example, Sturmfels and Sullivant (2005), Hendy et al. (1994), respectively].

The inspiration for the present paper can be seen as a generalization of the group-based models to ‘semigroup-based’ models. Below we show that this provides a previously unexplored method for deriving Lie-Markov models and we give a complete classification of all semigroup-based models in the cases of binary, three, and four (DNA) state models. In particular, we show that the well-known Felsenstein 81 model (Felsenstein 1981) can be interpreted as a semigroup-based model. This model is known to have rather tractable mathematical features leading to multiple analytic results [such as those given in Steel (2016, Chap 7)]. We argue that these results follow primarily from the underlying semigroup structure and similarly apply to any semigroup-based model. By way of illustration and with applications to identifying structural properties of molecular evolution in mind, in Sect. 4.2 we provide a detailed analysis of a novel semigroup-based model.

2 Rate Matrices, Stochasticity, and Lie-Markov Models

We consider homogeneous continuous-time Markov chains with state space $X = \{1, 2, \ldots, k\}$. A rate matrix $Q = (q_{ij})_{i, j \in X}$ is a $k \times k$ matrix with non-negative off-diagonal entries and zero column sums. Given a rate matrix $Q$, we may compute the corresponding probability transition matrix using the exponential map:

$$M(t) = e^{Qt},$$

where the $ji$ entry of $M(t)$ is the conditional probability $\mathbb{P}[i \to j, \text{ in time } t]$. 

\[ Springer \]
For our purposes, a *model* is, for fixed $k$, a choice of a restricted class of rate matrices $Q$. Taking the case of the DNA state space $X = \{A, G, C, T\} \equiv \{1, 2, 3, 4\}$ we have $k = 4$. The ever-popular general time-reversible model GTR (Tavaré 1986) is parametrized by the stationary distribution $\pi = (\pi_i)_{i \in X}$ of the chain and ‘relative’ rates $s_1, s_2, \ldots, s_6$:

$$Q_{\text{GTR}} = \begin{pmatrix} * & \pi_1 s_1 & \pi_1 s_2 & \pi_1 s_3 \\ \pi_2 s_1 & * & \pi_2 s_4 & \pi_2 s_5 \\ \pi_3 s_2 & \pi_3 s_4 & * & \pi_3 s_6 \\ \pi_4 s_3 & \pi_4 s_5 & \pi_4 s_6 & * \end{pmatrix},$$

(where the rows and columns are ordered according to $A, G, C, T$ and the missing entries $*$ are determined by the zero column-sum condition). The defining feature of this model is that it produces a Markov process that, at equilibrium, is identical to its time-reversed process. This property is exhibited by observing the matrix entries of this model that satisfy the detailed balance conditions $q_{ij} \pi_j = q_{ji} \pi_i$ for all choices $i \neq j$. The interpretation is that the rate of a transition $i \to j$ weighted by probability $\pi_i$ of being in state $i$ is equal to converse rate $j \to i$ weighted by probability $\pi_j$.

Another popular phylogenetic model is the HKY model (Hasegawa et al. 1985) which has the additional constraints $s_1 = s_6 = \kappa$ and $s_2 = s_3 = s_4 = s_5 = 1$, giving:

$$Q_{\text{HKY}} = \begin{pmatrix} * & \pi_1 \kappa & \pi_1 & \pi_1 \\ \pi_2 \kappa & * & \pi_2 & \pi_2 \\ \pi_3 & \pi_3 & * & \pi_3 \kappa \\ \pi_4 & \pi_4 & \pi_4 \kappa & * \end{pmatrix}.$$}

This model is again time-reversible and is motivated by distinguishing *transitions* (substitutions within purines and pyrimidines, i.e. $A \leftrightarrow G, C \leftrightarrow T$) from *transversions* (substitutions between purines and pyrimidines).

Below we also consider the symmetric model SYM which is obtained by taking $\pi$ to be the uniform distribution, and rescaling so that:

$$Q_{\text{SYM}} = \begin{pmatrix} * & s_1 & s_2 & s_3 \\ s_1 & * & s_4 & s_5 \\ s_2 & s_4 & * & s_6 \\ s_3 & s_5 & s_6 & * \end{pmatrix},$$

Roughly speaking, we say that a model is *multiplicatively closed* if, for any two substitution matrices $M_1$ and $M_2$ arising from the model, the matrix product $M_1 M_2$ is obtainable from the model. We will see below that the GTR, HKY, and SYM models are not multiplicatively closed.

An example of a well-known multiplicatively closed model is the sub-model of GTR obtained by setting each relative rate equal to unity (the Felsenstein 81 (Felsenstein}
or ‘equal-input’ (Steel 2016) model):

\[
Q_{F81} = \begin{pmatrix}
    * & \pi_1 & \pi_1 & \pi_1 \\
    \pi_2 & * & \pi_2 & \pi_2 \\
    \pi_3 & \pi_3 & * & \pi_3 \\
    \pi_4 & \pi_4 & \pi_4 & *
\end{pmatrix}.
\]

(1)

As was discussed in Sumner et al. (2012a), a sufficient condition for a multiplicatively closed model is that the set of rate matrices forms a Lie algebra. Under further reasonable assumptions regarding the construction of a model (essentially the model should be an intersection with an algebraic variety), the Lie algebra condition was more recently shown in Sumner (2017) to also be necessary. To state this result precisely, we follow the notation and definitions given in Sumner (2017).

For fixed \(k\), let \(L^+\) denote the set of \(k \times k\) rate matrices \(Q\). That is, \(Q \in L^+\) if \(Q\) has real, non-negative off-diagonal entries and zero column sums. We then let \(L \supset L^+\) denote the set of real \(k \times k\) matrices with zero column sums but with the non-negativity condition relaxed. A model \(R^+ \subseteq L^+\) is assumed to be expressible as an intersection \(R^+ = R \cap L^+\) where \(R \subseteq L\) is an algebraic variety (not necessarily irreducible). This means that \(R\) is defined by some polynomial constraints:

\[
R = \{ Q \in L : 0 = f_1(Q) = f_2(Q) = \cdots = f_r(Q) \},
\]

where each \(f_i(Q)\) is a polynomial in the entries of the matrix \(Q\). Further, we also assume that \(R\) is the minimal algebraic variety satisfying \(R^+ = R \cap L^+\).

We recall:

**Definition 2.1** A (matrix) Lie algebra is a set \(L\) of matrices satisfying, for all \(A, B \in L\) and scalars \(\lambda \in \mathbb{R}\), the two conditions:

(L1) \(A + \lambda B \in L\);
(L2) \([A, B] := AB - BA \in L\).

The first condition states that \(L\) forms a real vector space under sums and scalar multiplication. The operation \([A, B]\) is referred to as the ‘Lie bracket’ or ‘commutator’ and should be thought of as the natural product in the Lie algebra \(L\). (In the abstract formulation, a third condition known as the Jacobi identity is not needed here since we are restricting attention to matrix Lie algebras only and this condition is automatic in this case.)

**Definition 2.2** (Sumner 2017) A model \(R^+ = R \cap L^+\) is said to be multiplicatively closed if, for all choices \(Q_1, Q_2 \in R^+\) and \(t_1, t_2 \geq 0\), we have:

\[
\frac{1}{t_1 + t_2} \log(e^{Q_1 t_1} e^{Q_2 t_2}) \in R,
\]

where \(\log\) denotes the standard power series for the matrix logarithm.

As a consequence, we have:

\[
\hat{Q} := \frac{1}{t_1 + t_2} \log(e^{Q_1 t_1} e^{Q_2 t_2}) \implies e^{\hat{Q}(t_1 + t_2)} = e^{Q_1 t_1} e^{Q_2 t_2}.
\]
Thus, if $M_1 = e^{Q_{1t_1}}$ and $M_2 = e^{Q_{2t_2}}$ are derivable from the model, then so is their product $M_1 M_2$ (via the rate matrix $\hat{Q}$).

One should note that we intentionally do not insist on the stronger condition that $\hat{Q} \in \mathbb{R}^+$, since there are cases where $\log(e^{Q_{1t_1}} e^{Q_{2t_2}})$ has negative off-diagonal entries. The definition is designed such that taking products of substitution matrices and then the matrix log does not, from the geometric point of view, produce matrices outside of $\mathcal{R}$, which is, reassuringly, assumed to be minimal.

It follows that:

**Theorem 2.3** (Sumner 2017) A model $\mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+$ is multiplicatively closed if and only if $\mathcal{R}$ forms a Lie algebra.

For this reason, we refer to $\mathcal{R}^+$ as a Lie-Markov model whenever it is multiplicatively closed. The most immediate consequence of this result is that a multiplicatively closed model $\mathcal{R}^+$ must be determined by linear polynomial constraints $f_i(Q)$, that is, $\mathcal{R}$ is a linear space.

Comparative to the definition of a Lie algebra, consider:

**Definition 2.4** A matrix algebra is a set $\mathcal{A}$ of matrices satisfying, for all $A, B \in \mathcal{A}$ and scalars $\lambda$, the two conditions:

1. $A + \lambda B \in \mathcal{A}$;
2. $AB \in \mathcal{A}$.

As a simple consequence of these definitions:

**Lemma 2.5** Any matrix algebra $\mathcal{A}$ forms a (matrix) Lie algebra under commutators.

**Proof** We need only check that, for all $A, B \in \mathcal{A}$, we have $[A, B] \in \mathcal{A}$. But this follows easily since $AB \in \mathcal{A}$, $BA \in \mathcal{A}$, and $\mathcal{A}$ is closed under summation. □

There are certainly examples of (matrix) Lie algebras that do not form matrix algebras, so the converse is false in general. We give an example of a Lie-Markov model that does not form a matrix algebra in Sect. 3.3.

The GTR and HKY models are not Lie-Markov models since they are implicitly defined by nonlinear constraints on their matrix entries and hence fail to satisfy (L1). For the GTR model these constraints are known to be cubic by Kolmogorov’s criterion for detailed balance (Kolmogorov 1936). Similarly, the matrix entries of the HKY model satisfy the constraint $q_{12}q_{23} = q_{21}q_{13}$, and this constraint is not implied by simpler, linear conditions. On the other hand, the SYM and F81 models clearly satisfy (L1), but only F81 satisfies condition (L2) as well.

To see that the SYM model fails (L2), consider two rate matrices $Q_1, Q_2 \in \text{SYM}$ so $Q_i = Q_i^T$. Also assume $[Q_1, Q_2] \neq 0$ (such examples certainly exist). Now consider the transpose of the commutator:

$[Q_1, Q_2]^T = (Q_1 Q_2 - Q_2 Q_1)^T = Q_2^T Q_1^T - Q_1^T Q_2^T = [Q_2, Q_1] = -[Q_1, Q_2],$

so $[Q_1, Q_2]$ is an anti-symmetric matrix and (L2) fails to hold for this model.
To see that F81 model satisfies condition (L2), we exploit the fact that (L1) states that a Lie algebra is a vector subspace of matrices and hence has a basis. Confirming condition (L2) is then achieved by checking the commutators of all pairs of basis elements. We define the matrices \( \{ R_1, R_2, R_3, R_4 \} \) via

\[
Q = \begin{pmatrix}
* & \pi_1 & \pi_1 & \pi_1 \\
\pi_2 & * & \pi_2 & \pi_2 \\
\pi_3 & \pi_3 & * & \pi_3 \\
\pi_4 & \pi_4 & \pi_4 & *
\end{pmatrix} = \pi_1 R_1 + \pi_2 R_2 + \pi_3 R_3 + \pi_4 R_4 \in F81.
\]

(2)

Taking \( \mathcal{R}_{F81} = \text{span}_\mathbb{R}(R_1, R_2, R_3, R_4) \), explicit computation then shows

\[
[R_i, R_j] = R_i - R_j \in \mathcal{R}_{F81},
\]

so (L2) is satisfied, as required. Notice here that \([R_i, R_j]\) has some negative off-diagonal entries—this is why we are required to expand the definition of rate matrices to have entries from all of \( \mathbb{R} \). A major punchline for the approach we explore in this paper is that we will show how to derive the commutator relations (3) without the need to implement any matrix computations.

In the applied setting, we of course only use stochastic rate matrices. Thus, an additional feature of the theory is that, given a Lie-Markov model \( \mathcal{R}^+ \) is defined as the intersection \( \mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+ \), in general there are multiple approaches to parametrizing \( \mathcal{R}^+ \) as a subset of \( \mathcal{L} \). General tools for finding sensible parametrizations are discussed in Fernández-Sánchez et al. (2015) and Woodhams et al. (2015).

### 2.1 Model Symmetries

Suppose \( \mathcal{R}^+ \subset \mathcal{L}^+ \) is a Markov model on \( k \)-states (not necessarily multiplicatively closed) and \( G \leq S_k \) is a permutation subgroup of the symmetric group \( S_k \).

**Definition 2.6** We say that \( \mathcal{R}^+ \) has \( G \)-symmetry if, for all \( \sigma \in G \), we have:

\[
Q \in \mathcal{R}^+ \implies K_\sigma Q K_\sigma^T \in \mathcal{R}^+,
\]

where \( K_\sigma \) is the standard \( k \times k \) permutation matrix corresponding to \( \sigma \).

In other words if, according to the permutation \( \sigma \in G \), we simultaneously permute the rows and columns of a rate matrix in the model we obtain another rate matrix also in the model.

It is not hard to see that the GTR, SYM, and F81 models have \( S_4 \) symmetry, whereas the HKY model has reduced symmetries given by the dihedral group:

\[
D_4 = \{ e, (12), (34), (12)(34), (13)(24), (14)(23), (1324), (1423) \}.
\]

In Sect. 3.4, we will discuss the stronger notion of model symmetry used to define the equivariant models.
We also use this notion to define model equivalence:

**Definition 2.7** Two models \( R^+_1, R^+_2 \in L^+ \) on \( k \)-states are **isomorphic** if there exists a permutation \( \sigma \in S_k \) such that:

\[
Q \in R^+_1 \iff K_\sigma Q K^T_\sigma \in R^+_2.
\]

Considering the HKY model, given that the eight permutations \( \sigma \in D_4 \) are the only symmetries of this model, we see that there should exist \( 4! / 8 = 3 \) isomorphic variants of this model. These are easily understood as corresponding to the three possible partitionings of nucleotides into two sets of size two: \( AG|CT \), \( AT|CG \), and \( AC|GT \).

In Sumner et al. (2012a), we enumerated all the Lie-Markov DNA models with full symmetry \( S_4 \) and followed this up in Fernández-Sánchez et al. (2015) by enumerating all the Lie-Markov models with dihedral symmetry \( D_4 \). In Woodhams et al. (2015) we then explored the performance of these models on real data sets taking note of the three hierarchies of models that, similarly to the HKY model, arise from the three choices of nucleotide partitions.

### 3 Semigroup-Based Models

We begin by showing how to interpret the F81 model (1) as arising from an (abstract) semigroup of degree four. We then show that this semigroup generalizes to degree \( k \) and we similarly obtain the ‘equal-input’ model (Steel 2016) on any number of states \( k \). We then set up a general framework for deriving semigroup-based models and explain how this is a natural generalization of the notion of a group-based model.

**Definition 3.1** A **semigroup** is a set \( S \) equipped with an associative binary operation. In the finite case, we refer to \( |S| \) as the *degree* of \( S \).

Throughout, we denote the binary operation on a semigroup \( S \) multiplicatively (using concatenation), so, for all \( a_1, a_2, a_3 \in S \) we have \( a_1(a_2a_3) = (a_1a_2)a_3 \). We note that a group is a semigroup equipped with an identity element \( e \) satisfying \( ae = ea = a \) and inverses \( a^{-1} \) satisfying \( a^{-1}a = e = aa^{-1} \), for all \( a \in S \).

Consider the semigroup \( S = \{a_1, a_2, a_3, a_4\} \) with, for all \( i, j \in \{1, 2, 3, 4\} \), multiplication given by \( a_i a_j = a_{ij} \). To confirm this is indeed semigroup, we need only check that multiplication is associative: \( (a_ia_j)a_k = a_ia_(jk) = a_i = a_ia_j = a_i(a_ja_k) \).

We may represent this semigroup using \( 4 \times 4 \) matrices \( \{A_1, A_2, A_3, A_4\} \) defined by their action on standard unit vectors \( e_i \) via \( A_ie_j = e_i \). As a consequence we have \( A_i A_j e_k = A_i e_j = e_i \) and hence \( A_i A_j = A_i \), which shows that these matrices mimic the multiplication rule \( a_i a_j = a_i \) in general. We also see that \( A_i \) is the matrix with 1s on row \( i \) and zeros elsewhere. For example,

\[
A_1 = \begin{pmatrix}
1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}.
\]
We then construct the rate matrices \( R_i := -I + A_i \in \mathcal{L}^+ \) and see that these are none other than the basis elements for the F81 model given in (2), for example:

\[
R_1 = \begin{pmatrix}
0 & 1 & 1 & 1 \\
0 & -1 & 0 & 0 \\
0 & 0 & -1 & 0 \\
0 & 0 & 0 & -1
\end{pmatrix} \in \mathcal{R}_{\text{F81}}^+.
\]

Further, as promised in the previous section, the commutators follow immediately from the multiplication rule in the semigroup:

\[
\left[ R_i, R_j \right] = [A_i, A_j] = A_i A_j - A_j A_i = A_i - A_j = R_i - R_j,
\]

where, in the first equality, we have used the fact that commutators are blind to inclusion of scalar multiples of the identity matrix \( I \).

This example immediately generalizes as follows.

**Definition 3.2** The left zero semigroup of degree \( k \) is the semigroup \( \mathcal{B}_k = \{a_1, a_2, \ldots, a_k\} \) with multiplication rule \( a_i a_j = a_i \).

Applying the above procedure to \( \mathcal{B}_k \), we obtain what is referred to as the ‘equal-input’ model (Steel 2016) with rate matrices:

\[
Q = \alpha_1 R_1 + \alpha_2 R_2 + \cdots + \alpha_k R_k = \begin{pmatrix}
* & \alpha_1 & \alpha_1 & \ldots & \alpha_1 \\
\alpha_2 & * & \alpha_2 & \ldots & \alpha_2 \\
& \vdots & & \ddots & \\
\alpha_k & \alpha_k & \alpha_k & \ldots & *
\end{pmatrix},
\]

and commutators \( \left[ R_i, R_j \right] = R_i - R_j \). We will see that the left zero semigroup plays a pivotal role in Result 1.

This procedure generalizes to produce a \( k \)-state Lie-Markov model from any degree \( k \) semigroup. Our process of converting each semigroup element \( a_i \) into a matrix is inspired from taking the ‘regular representation’ of a group. In detail, suppose \( a_i, a_j, a_k \in S \) satisfy \( a_k = a_i a_j \). Then the \( j \)th column of the matrix \( A_i \) has a single nonzero entry 1 in the \( k \)th row. It is then clear that \( A_k = A_i A_j \), mimicking the multiplication in the semigroup, and each \( L_i := -I + A_i \in \mathcal{L}^+ \) is a rate matrix. It is important to note, however, that the resulting map from the semigroup to the matrices \( A_i \) is not necessarily injective (as it is for a group).

Given an enumeration of all semigroups of size \( k \), our general procedure for producing semigroup-based models with \( k \)-states is then, for each semigroup \( S = \{a_1, a_2, \ldots, a_k\} \):

1. List the set of \( k \times k \) matrices \( A_1, A_2, \ldots, A_k \) resulting from the regular representation of \( S \);
2. Define the rate matrices \( L_i = -I + A_i \in \mathcal{L}^+ \);
3. Take \( \mathcal{R} = \text{span}_{\mathbb{R}} (L_1, L_2, \ldots, L_k) \) and define the model \( \mathcal{R}^+ = \mathcal{R} \cap \mathcal{L}^+ \).
We discuss the connection of semigroup-based models to the usual construction of group-based models in the next section. Presently, we note:

**Theorem 3.3** Every semigroup-based model is a Lie-Markov model.

**Proof** (L1) is true by construction of \( R \) as a linear span, and (L2) follows by the multiplicative closure in the semigroup \( S \). Indeed, suppose \( a_i, a_j \in S \) satisfy \( a_i a_j = a_k \), then:

\[
L_i L_j = (-I + A_i)(-I + A_j) = I - A_i - A_j + A_k = -L_i - L_j + L_k \in R,
\]

so, applying Lemma 2.5, we see that \( R \) forms a Lie algebra. \( \square \)

The reader should note that, since the regular representation of a semigroup is not necessarily injective, the matrices \( A_1, A_2, \ldots, A_k \) need not be distinct. Further, as we will see in Sect. 3.1, it is also possible that some \( A_i = I \) and consequently \( L_i = 0 \). Thus, in general \( \dim(R) \leq |S| \). This means that there are examples of non-isomorphic semigroups that produce isomorphic (or even equal) Markov models. An example is given in Sect. 4.3.

We recall that two semigroups \( S, S' \) are **isomorphic** if there exists a bijection \( \varphi : S \rightarrow S' \) satisfying, for all \( s_1, s_2 \in S \):

\[
\varphi(s_1 s_2) = \varphi(s_1) \varphi(s_2).
\]

It should be clear that the Markov models produced by two isomorphic semigroups differ only up to a possible permutation of states. Thus, in what follows, we only need to consider non-isomorphic semigroups.

We also recall that two semigroups \( S, S' \) are **anti-isomorphic** if there exists a bijection \( \overline{\varphi} : S \rightarrow S' \) satisfying, for all \( s_1, s_2 \in S \):

\[
\overline{\varphi}(s_1 s_2) = \overline{\varphi}(s_2) \overline{\varphi}(s_1).
\]

An easy way to construct an anti-isomorphism is to take a semigroup \( S \) and then define \( S' \) by reversing the multiplication on \( S \):

\[
s_1 s_2 = s_3 \text{ in } S \iff s_2 s_1 = s_3 \text{ in } S'.
\]

In this case it is said that \( S' \) and \( S \) are **dual semigroups**.

When enumerating semigroups it is often the case that dual semigroups are not treated separately. However, for our purposes, it is the case that the two semigroups can produce radically different Markov models. This phenomenon is illustrated with the following example.

### 3.1 Dual Semigroups Can Produce Different Lie-Markov Models

Consider the dual of the left zero semigroup \( B_4 \) underlying the F81 model above: \( B'_4 = \{ a_1, a_2, a_3, a_4 \} \) with multiplication rule \( a_i a_j = a_j \). The regular representation...
produces the matrices

$$A_1 = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix} = I = A_2 = A_3 = A_4,$$

which (beyond being non-injective) gives the trivial Lie-Markov model where all rate matrices are zero: $Q = 0$.

Thus, we treat dual semigroups independently in what follows.

### 3.2 Different Lie-Markov Models Can form Isomorphic Lie Algebras

Consider the general two-state Markov model:

$$Q = \begin{pmatrix}
-\alpha & \beta \\
\alpha & -\beta
\end{pmatrix} = \alpha \begin{pmatrix}
-1 & 0 \\
0 & -1
\end{pmatrix} + \beta \begin{pmatrix}
0 & 1 \\
0 & -1
\end{pmatrix} = \alpha L_1 + \beta L_2.$$

This is a Lie-Markov model since it forms a matrix algebra:

$$L_1^2 = -L_1, \quad L_1 L_2 = -L_2, \quad L_2 L_1 = -L_1, \quad L_2^2 = -L_2,$$

and hence, following Lemma 2.5, a Lie algebra:

$$[L_1, L_2] = L_1 - L_2.$$

The geometric consequences of the identification of the Lie algebra underlying this model are explored in Sumner (2013).

We will not review the construction, but an interesting three-state model arises from the two-state general Markov model using the method given in Jarvis and Sumner (2012). This model has rate matrices given by

$$Q' = \begin{pmatrix}
-2\alpha & \beta & 0 \\
2\alpha & -\alpha - \beta & 2\beta \\
0 & \alpha & -2\beta
\end{pmatrix} = \alpha \begin{pmatrix}
-2 & 0 & 0 \\
2 & -1 & 0 \\
0 & 1 & 0
\end{pmatrix} + \beta \begin{pmatrix}
0 & 1 & 0 \\
0 & -1 & 2 \\
0 & 0 & -2
\end{pmatrix} = \alpha L'_1 + \beta L'_2$$

satisfying

$$[L'_1, L'_2] = L'_1 - L'_2.$$

Since these two models define two-dimensional Lie algebras satisfying the same commutator relations, they are isomorphic as Lie algebras. In fact, for any number of character states $k$, the method given in Jarvis and Sumner (2012) produces a two-dimensional Lie-Markov model which is isomorphic, as a Lie algebra, to the two-state general Markov model. This illustrates that Lie algebra isomorphism is not (in itself) a useful tool for identifying distinct Lie-Markov models.
3.3 Not All Lie-Markov Models form Matrix Algebras

For completeness of discussion, we give an example of a Lie-Markov model that forms a Lie algebra (Definition 2.1) without satisfying the stronger condition of forming a matrix algebra (Definition 2.4). In fact, the three-state example from the previous section is sufficient.

Consider the matrix product:

\[
L_1' L_2' = \begin{pmatrix}
0 & -2 & 0 \\
0 & 3 & -2 \\
0 & -1 & 2
\end{pmatrix}.
\]

This is clearly not expressible as a linear combination of \(L_1'\) and \(L_2'\). Thus, we see that there exist examples of Lie-Markov models which do not form matrix algebras.

3.4 Equivariant Models

As mentioned in the introduction, an important class of Markov models that form matrix algebras (and hence are Lie-Markov models) are the ‘equivariant’ models (Draisma and Kuttler 2009). These models were originally defined as sets of substitution matrices, but the idea is easily translated into the setting of rate matrices, as was described in Sumner et al. (2012a) and reproduced presently.

Fix a permutation group \(G \leq S_k\). The equivariant model corresponding to \(G\) is then obtained by taking the set of rate matrices \(Q\) which are invariant under simultaneous row and column permutations by \(\sigma \in G\). Concretely, if \(\sigma \in G\) is a permutation and \(K_\sigma\) is the corresponding \(k \times k\) permutation matrix, then the rate matrices \(Q\) in the equivariant model have the defining feature

\[
K_\sigma Q K_\sigma^T = Q.
\]

A simple consequence is that each equivariant model forms a matrix algebra via

\[
K_\sigma (Q_1 Q_2) K_\sigma^T = (K_\sigma Q_1 K_\sigma^T)(K_\sigma Q_2 K_\sigma^T) = Q_1 Q_2.
\]

Hence, following Lemma 2.5, the equivariant models form Lie algebras and are therefore Lie-Markov models.

For example, if we take the group of dihedral permutations

\[
D_4 = \{e, (12), (34), (12)(34), (13)(24), (14)(23), (1324), (1423)\}
\]
(the symmetries of a square), we obtain the Kimura two-parameter model (Kimura 1980) as an equivariant model:

\[ Q_{K2ST} = \begin{pmatrix}
* & \alpha & \beta & \beta \\
\alpha & * & \beta & \beta \\
\beta & \beta & * & \alpha \\
\beta & \beta & \alpha & *
\end{pmatrix}. \tag{4} \]

### 3.5 Connection to Group-Based Models

Here we describe the usual construction of group-based models (Semple and Steel 2003). We recall that the construction of a group-based model is usually thought of being valid only for abelian (finite) groups. However, following Sumner et al. (2012a), we reinterpret the construction using the concept of the regular representation of a group and show how this allows us to construct models for general, possibly non-abelian (finite) groups. As in the case of the equivariant models, the construction can be implemented using a substitution or rate matrix formulation and the results are equivalent.

Given a finite abelian group \( G \), fix a linear function \( f : G \to \mathbb{R}_{\geq 0} \) denoted as \( f(g) = \alpha_g \) for all \( g \in G \). Using additive notation in \( G \), for each pair \( i, j \in G \) let the rate of substitution \( i \to j \) be given by \( f(i - j) \) where \( i - j \in G \). Then we construct the rate matrix \( Q \) with off-diagonal entries \( q_{ij} = f(i - j) \) and diagonal entries determined by the zero sum condition.

Repeating what is done in Sumner et al. (2012a), we show that this concept naturally extends to a general finite group \( G \) (not necessarily abelian) by invoking the concept of the regular representation. We recall that the regular representation of \( G \) is given by mapping each \( g \in G \) to a \(|G| \times |G| \) (permutation) matrix \( K(g) \) by setting the entry corresponding to each pair \( g_1, g_2 \in G \) equal to 1 if \( g_1 = gg_2 \) and equal to 0 otherwise. These matrices then satisfy the rule \( K(g)K(g') = K(gg') \) and, as above, we may define the rate matrices

\[ L_g = -I + K(g), \]

which naturally form a matrix (and hence Lie) algebra:

\[ L_gL_{g'} = I - K(g) - K(g') + K(gg') = -L_g - L_{g'} + L_{gg'}. \]

If \( G \) is abelian, it is not hard to show we obtain exactly the group-based model corresponding to \( G \). However, under this construction it is no longer necessary for \( G \) to be abelian.

For example, consider the Klein 4-group:

\[ V_4 = \{e, (12)(34), (13)(24), (14)(23)\} \cong C_2 \times C_2, \]
where $C_2$ is the cyclic group on two elements $C_2 \cong \mathbb{Z}_2$. This group produces the well-known Kimura 3ST model (Kimura 1981):

$$
\begin{pmatrix}
\ast & \alpha & \beta & \delta \\
\alpha & \ast & \delta & \beta \\
\beta & \delta & \ast & \alpha \\
\delta & \beta & \alpha & \ast
\end{pmatrix}.
$$

(5)

As the simplest non-abelian example, we set $G = S_3$, the symmetric group on three elements, and obtain the six-state Markov model with rate matrices:

$$Q = \alpha_{(12)} L_{(12)} + \cdots + \alpha_{(132)} L_{(132)} =
\begin{pmatrix}
\ast & \alpha_{(12)} & \alpha_{(132)} & \alpha_{(123)} & \alpha_{(23)} & \alpha_{(13)} \\
\alpha_{(12)} & \ast & \alpha_{(23)} & \alpha_{(13)} & \alpha_{(132)} & \alpha_{(123)} \\
\alpha_{(123)} & \alpha_{(23)} & \ast & \alpha_{(132)} & \alpha_{(13)} & \alpha_{(12)} \\
\alpha_{(132)} & \alpha_{(13)} & \alpha_{(123)} & \ast & \alpha_{(12)} & \alpha_{(23)} \\
\alpha_{(23)} & \alpha_{(123)} & \alpha_{(13)} & \alpha_{(12)} & \ast & \alpha_{(132)} \\
\alpha_{(13)} & \alpha_{(132)} & \alpha_{(12)} & \alpha_{(23)} & \alpha_{(123)} & \ast
\end{pmatrix},
$$

and commutators

$$[L_\sigma, L_{\sigma'}] = L_{\sigma \sigma'} - L_{\sigma' \sigma},$$

for all choices $\sigma, \sigma' \in S_3$.

Although this generalization to non-abelian groups is mathematically appealing, it is not of much use in phylogenetics with $k = 4$ DNA states, since there are exactly two groups of degree 4: $C_4$ and $C_2 \times C_2$, both of which are abelian, and hence are already obtainable using the standard approach to group-based models. However, the astute reader will have noticed that there is nothing in the above that uses the availability of algebraic inverses in the group $G$. Hence, we may generalize immediately to semigroups $S$ and obtain exactly the construction given above for deriving semigroup-based models.

4 Results

In this section we present our results of exploring the semigroup-based models derived from semigroups of degree $k = 2$, 3, and 4.

To begin, we recall:

**Definition 4.1** Consider a Markov chain on a state space $X$.

(i). The chain is **irreducible** if, for all pairs $i, j \in X$, it is possible to transition from $i$ to $j$ in a finite number of steps. A Markov chain that is not irreducible is said to be **reducible**.

(ii). A state $i \in X$ is an **absorbing state** if it is impossible to transition from $i$ to any other state.
We note that a Markov chain with an absorbing state is clearly reducible. We consider models that form reducible Markov chains to not be of interest for our motivations in phylogenetic modelling.

Following Forsythe (1955), there are 188 non-isomorphic semigroups of degree 4 with 126 of these being neither isomorphic nor anti-isomorphic (in other words there are 62 semigroups that occur as dual pairs). We implemented our procedure for deriving semigroup-based models on all 188 of these semigroups and then removed those models which are the same under permutation of nucleotide states (c.f. Definition 2.7). From this process we produced 114 non-isomorphic Lie-Markov models. We repeated this procedure on the 5 non-isomorphic semigroups of degree 2 and then on the 24 non-isomorphic semigroups of degree 3. A complete list of the semigroup-based models we produced is given in the supplementary material.

However, the majority of models we obtained from this procedure form reducible Markov chains. After removal of the reducible models, we found:

**Result 1** For \( k = 2, 3, \) and 4 states, the non-isomorphic, irreducible semigroup-based models are precisely those arising from the semigroups that occur as direct products of cyclic groups and left zero semigroups.

Complete information is provided in Table 1. Examples of semigroup-based models which are reducible are given in the two- and three-state results below.

### 4.1 Four-State Semigroup Models

The two four-state models not yet discussed are as follows.

Model 3.3b: We recall that the rate matrices in Model 3.3b (Fernández-Sánchez et al. 2015) are expressible in the form

\[
Q_{3.3b} = \begin{pmatrix}
* & \alpha & \gamma & \beta \\
\alpha & * & \gamma & \beta \\
\beta & \gamma & * & \alpha \\
\gamma & \beta & \alpha & *
\end{pmatrix},
\]

which can be understood as the ‘twisted’ cousin to the Kimura 3ST model (5). It is amusing to ‘reverse-engineer’ this model to find it is based on the cyclic group \( \mathbb{C}_4 \) with multiplication table:

|   | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( a_4 \) |
|---|---|---|---|---|
| \( a_1 \) | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( a_4 \) |
| \( a_2 \) | \( a_2 \) | \( a_1 \) | \( a_4 \) | \( a_3 \) |
| \( a_3 \) | \( a_3 \) | \( a_4 \) | \( a_2 \) | \( a_1 \) |
| \( a_4 \) | \( a_4 \) | \( a_3 \) | \( a_1 \) | \( a_2 \) |

Since \( \mathbb{C}_4 \) is an abelian group, the associated Lie algebra is also abelian and hence all commutators are trivial.

Following Fernández-Sánchez et al. (2015), this model has the symmetries of the dihedral group \( \mathbb{D}_4 \) and hence respects the partitioning of nucleotides into purine and...
Table 1  Classification of irreducible semigroup-based models for \( k = 2, 3, \) and 4 states

| Semigroup-based model        | dim | Semigroup | symm |
|------------------------------|-----|-----------|------|
| 4 states (188)               |     |           |      |
| F81/equal-input/4.4a         | 4   | \( B_2 \times B_2 \) | \( S_4 \) |
| K3ST/3.3a                    | 3   | \( ^* C_2 \times C_2 \) | \( S_4 \) |
| ‘Twisted’ K3ST/3.3b          | 3   | \( ^* C_4 \) | \( D_4 \) |
| New model                    | 4   | \( C_2 \times B_2 \) | \( V_4 \) |
| 3 states (24)                |     |           |      |
| Equal-input                  | 3   | \( B_3 \) | \( S_3 \) |
| Group-based                  | 2   | \( ^* C_3 \) | \( S_3 \) |
| 2 states (5)                 |     |           |      |
| General Markov/equal-input   | 2   | \( B_2 \) | \( S_2 \) |
| Binary symmetric             | 1   | \( ^* C_2 \) | \( S_2 \) |

The Lie-Markov model identifier (e.g. ‘4.4a’) follows the notation used in Fernández-Sánchez et al. (2015) and \(^*\) indicates the semigroup is a group (and hence the model is ‘group-based’). In each case, the number of non-isomorphic semigroups is indicated in parentheses. The second column indicates the dimension of the model as a Lie algebra and corresponds to the number of free rate parameters. The fourth column indicates the model symmetry (according to Definition 2.6). The reader should note that there is only one degree 4 left zero semigroup (i.e. \( B_4 \cong B_2 \times B_2 \)), so the enumeration in the four-state case is complete (c.f. Result 1).

Pyrimidines. This also tells us that there are 3 = \( 4! / 8 \) distinct (but isomorphic) copies of this model corresponding to the three nucleotide partitionings: \( AG|CT, AT|CG, \) and \( AC|GT \).

\( C_2 \times B_2 \) model: The model previously unknown to us is based on the semigroup \( C_2 \times B_2 \) with multiplication table

|     | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( a_4 \) |
|-----|----------|----------|----------|----------|
| \( a_1 \) | \( a_1 \) | \( a_1 \) | \( a_3 \) | \( a_3 \) |
| \( a_2 \) | \( a_2 \) | \( a_2 \) | \( a_4 \) | \( a_4 \) |
| \( a_3 \) | \( a_3 \) | \( a_3 \) | \( a_1 \) | \( a_1 \) |
| \( a_4 \) | \( a_4 \) | \( a_4 \) | \( a_2 \) | \( a_2 \) |

and rate matrices of the form

\[
Q = \begin{pmatrix}
\ast & \alpha & \gamma & \gamma \\
\beta & \ast & \delta & \delta \\
\gamma & \gamma & \ast & \alpha \\
\delta & \delta & \beta & \ast
\end{pmatrix} = \alpha L_1 + \beta L_2 + \gamma L_3 + \delta L_4,
\]

(6)

with commutators

\[
[L_1, L_2] = L_1 - L_2, \quad [L_1, L_3] = 0, \quad [L_1, L_4] = L_3 - L_4,
\]

\[
[L_2, L_3] = L_4 - L_3, \quad [L_2, L_4] = 0, \quad [L_3, L_4] = L_1 - L_2,
\]
Direct computation shows that the symmetry group of this model is the Klein 4-group:

\[ V_4 = \{e, (12)(34), (13)(24), (14)(23)\}. \]

This means there are \( 6 = 4!/4 \) distinct (but isomorphic) variants of this model to test on real data sets. We describe these variants more completely presently.

### 4.2 Analysis of the \( C_2 \times B_2 \) Model

We simplify notation and set \( C_2 = \{e, \sigma\} \) with \( \sigma^2 = e \), and \( B_2 = \{a, b\} \) with \( a^2 = ab = a \) and \( b^2 = ba = b \). Referring to (6), we have a correspondence between the semigroup elements \( (e, a), (e, b), (\sigma, a), (\sigma, b) \in C_2 \times B_2 \) and the substitution rates \( \alpha, \beta, \gamma, \delta \), respectively.

We recall that nucleotides can be partitioned into pairs in three distinct ways: \( AG|CT \) corresponding to division into purines/pyrimidines (two/one nitrogen rings, respectively), \( AT|CG \) motivated from the ‘Watson-Crick’ base pairing that occurs across the complementary stands of the DNA helix, and \( AC|GT \) corresponding to amino/keto forms, respectively. The six variants of the \( C_2 \times B_2 \) model are then obtained by (i) choosing two out of the three available partitionings and (ii) assigning substitutions within the blocks of the first partition to \( C_2 \) and substitutions within the blocks of the second partition to \( B_2 \).

Many well-known phylogenetic models are constructed to explicitly respect either the partitioning into purine/pyrimidines [e.g. (4); the Kimura two-parameter model (Kimura 1980)] or into Watson–Crick pairs (Yap and Pachter 2004). Thus, the most immediately biologically compelling variants of the \( C_2 \times B_2 \) model are obtained by choosing the purine/pyrimidine (RY) and Watson–Crick (WC) partitionings and assigning them to \( C_2 \) and \( B_2 \) in the two available ways. In this way we obtain two distinct models, which we presently describe explicitly.

To begin, we note that we may appeal to the \( V_4 \) symmetries of the model to justify canonically choosing an encoding with \( (e, a) \equiv A \) in each case. The ‘\( C_2^{RY} \times B_2^{WC} \)’ model then follows from the encoding:

\[
(e, a) \equiv A \quad (e, b) \equiv T \quad (\sigma, a) \equiv G \quad (\sigma, b) \equiv C
\]

thereby producing the substitution rates

\[
\begin{align*}
\alpha &: T \to A & \beta &: A \to T & \gamma &: C \to A & \delta &: G \to T.
\end{align*}
\]

Correspondingly, the ‘\( C_2^{WC} \times B_2^{RY} \)’ model follows from the encoding:

\[
(e, a) \equiv A \quad (e, b) \equiv G \quad (\sigma, a) \equiv T \quad (\sigma, b) \equiv C
\]
with substitution rates

\[
\begin{align*}
\alpha &: \ G \rightarrow A \quad C \rightarrow T \\
\beta &: \ A \rightarrow G \quad T \rightarrow C \\
\gamma &: \ C \rightarrow A \quad G \rightarrow T \\
\delta &: \ A \leftrightarrow T \quad G \leftrightarrow C
\end{align*}
\]

Clearly, both of these variants possess interesting, but distinct, biological structure relative to the RY and WC pairings. In particular, we note that the \(\mathbf{C}_2^{\text{RY}} \times \mathbf{B}_2^{\text{WC}}\) model treats within purine and within pyrimidine substitutions symmetrically, while the \(\mathbf{C}_2^{\text{WC}} \times \mathbf{B}_2^{\text{RY}}\) model is ‘strand-symmetric’ (Casanellas and Sullivant 2005). Further, if we switch the two models, these statements become false. It is an interesting topic for future work to explore whether there are examples of molecular data sets that have a particular preference for either model and to see what can be learnt from such a comparison.

We continue our analysis of the \(\mathbf{C}_2 \times \mathbf{B}_2\) model by observing that (via a similarity transformation) we can bring the model into a block diagonal form which allows explicit formulas for the associated substitution probabilities to be calculated. Our inspiration to do this follows the process of Hadamard conjugation or Fourier analysis of general group-based models (see Steel (2016, Chap. 7.4), for a recent review).

In the next section we will see that the group \(\mathbf{C}_2\) produces the group-based model usually referred to as the ‘binary symmetric’ model, which, as is well known, is diagonalized by the Hadamard matrix \(h = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\). Similarly, we will see that the semigroup \(\mathbf{B}_2\) produces the general two-state Markov model (or, equivalently, the two-state equal-input model). The general two-state Markov model does not admit a diagonalization, however, as is well known, the associated substitution matrices admit a simple analytic form which enables further analysis (see, for example, Kingman (1962) for the solution of the ‘embedding’ problem and Sumner (2013) for decomposition in terms of ideals of the underlying Lie algebra). This inspires us to explore the \(\mathbf{C}_2 \times \mathbf{B}_2\) model using the similarity transformation obtained from the Kronecker product:

\[
J = h \otimes I = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.
\]

Indeed, applied to the rate matrices (6), explicit computation produces the block form:

\[
J^{-1} Q J = \begin{pmatrix} -\beta - \delta & \alpha + \gamma & 0 & 0 \\ \beta + \delta & -\alpha - \gamma & 0 & 0 \\ 0 & 0 & -\beta - 2\gamma - \delta & \alpha - \gamma \\ 0 & 0 & \beta - \delta & -\alpha - \gamma - 2\delta \end{pmatrix}.
\]

This is analogous to the block forms for the strand-symmetric model derived using more general arguments in Casanellas and Sullivant (2005) and Jarvis and Sumner (2016), which, in those papers, were then applied to the exploration of phylogenetic
invariants and Markov invariants for the strand-symmetric model, respectively. Of course, the above form does not generally give a rate matrix. However, we observe that the upper $2 \times 2$ block is a rate matrix, which, assuming the $\mathbf{C}_2^{\text{RY}} \times \mathbf{B}_2^{\text{WC}}$ encoding, explicitly shows that we may ‘clump’ the purines into one state and pyrimidines into another state thereby producing a two-state Markov chain with purine to pyrimidine rate $\alpha + \gamma$ and pyrimidine to purine rate $\beta + \delta$.

If we define the parameters $x = \alpha + \gamma$, $y = \beta + \delta$, $u = \alpha - \gamma$, and $v = \beta - \delta$, we find:

$$J^{-1}QJ = \begin{pmatrix}
-y & x & 0 & 0 \\
-\alpha & y & 0 & 0 \\
0 & 0 & -k + u & u \\
0 & 0 & v & -k + v
\end{pmatrix},$$

where $k = x + y = \alpha + \beta + \gamma + \delta$ If we normalize the sum of rates to $k = x + y = 1$, a straightforward computation with the exponential power series allows us to write, for all $t \geq 0$:

$$J^{-1}e^{Qt}J = \begin{pmatrix}
\begin{array}{cccc}
e^{-t}y + x & x(1 - e^{-t}) & 0 & 0 \\
\alpha e^{t(u+v)} & 0 & 0 & 0 \\
\alpha e^{t(u+v)} & 0 & 0 & 0 \\
0 & 0 & \frac{ue^{t(u+v) + v}}{u+v} & \frac{ue^{t(u+v) + v}}{u+v}
\end{array}
\end{pmatrix}.$$

This shows that the substitution probabilities for this model can be expressed in relatively simple analytic form; in analogy to the substitution matrices for the equal-input model [as discussed in Steel (2016, Chap. 7.3)].

We leave further analysis and interpretation of this model to future work.

### 4.3 Two-State Semigroup Models

In the case of $k = 2$ binary states, there are precisely five non-isomorphic semigroups. We denote each as $S = \{a_1, a_2\}$.

Semigroup 1:

\[
\begin{array}{cccc}
a_1 & a_2 \\
a_1 & a_1 \\
a_2 & a_1 \\
\end{array}
\]

Taking the regular representation produces the matrices

\[
A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.
\]
Thus, the regular representation is not injective and produces the one-dimensional Lie-Markov model with an absorbing state and rate matrices given by:

\[ Q = \begin{pmatrix} 0 & \alpha \\ 0 & -\alpha \end{pmatrix}. \]

Since this model treats the two states very differently, it has trivial symmetry group \( \{e\} < S_2 \).

Semigroup 2:

\[
\begin{array}{c|cc}
& a_1 & a_2 \\
\hline
a_1 & a_1 & a_1 \\
a_2 & a_1 & a_2 \\
\end{array}
\]

Taking the regular representation produces the matrices

\[ A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \]

This time the representation is injective, but, since \(-I + A_2 = 0\), this semigroup produces exactly the same Lie-Markov model as the previous case.

Semigroup 3:

\[
\begin{array}{c|cc}
& a_1 & a_2 \\
\hline
a_1 & a_1 & a_1 \\
a_2 & a_2 & a_2 \\
\end{array}
\]

Taking the regular representation produces the matrices

\[ A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}. \]

The resulting Markov model can equivalently be understood as producing the two-state equal-input model or the two-state general Markov model (as discussed in Sects. 3 and 3.2, respectively). This model has full symmetry \( S_2 \).

Semigroup 4:

\[
\begin{array}{c|cc}
& a_1 & a_2 \\
\hline
a_1 & a_1 & a_2 \\
a_2 & a_1 & a_2 \\
\end{array}
\]

This is the dual semigroup to the previous case.

Taking the regular representation produces the matrices

\[ A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \]
which is not injective and produces the trivial Lie-Markov model with vanishing rate
matrices $Q = 0$.

Semigroup 5:

\[ \begin{array}{ccc}
   a_1 & a_2 \\
   a_1 & a_2 \\
   a_2 & a_1 \\
\end{array} \]

This is the (abelian) cyclic group on two elements $C_2 \cong \mathbb{Z}_2$. Taking the regular
representation produces the matrices

\[ A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \]

which produces the group-based model

\[ Q = \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix}, \]

also known as the ‘binary symmetric’ model. This is a one-dimensional Lie-Markov
model with trivial commutator relations. This model has full symmetry $S_2$.

### 4.4 Three-State Semigroup Models

There are 24 non-isomorphic semigroups with degree 3. The 24 cases split into 12
which are ‘self’ dual plus 6 anti-isomorphic pairs. From these, we found that there are
15 non-isomorphic semigroup-based models, two of which form irreducible Markov
chains. We provide a complete set of results in the supplementary material and simply
present the interesting cases here.

Example 1:

\[ \begin{array}{ccc}
   a_1 & a_2 & a_3 \\
   a_1 & a_1 & a_1 \\
   a_2 & a_2 & a_2 \\
   a_3 & a_3 & a_3 \\
\end{array} \]

Taking the regular representation produces the matrices

\[ A_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \]

which produces the three-state equal-input model:

\[ Q = \alpha R_1 + \beta R_2 + \gamma R_3 = \begin{pmatrix} -\beta - \gamma & \alpha & \alpha \\ \beta & -\alpha - \gamma & \beta \\ \gamma & \gamma & -\alpha - \beta \end{pmatrix}, \]

\[ \text{Springer} \]
forming a matrix (and hence Lie) algebra with relations:

\[ R_i R_j = -R_j \implies [R_i, R_j] = R_i - R_j. \]

This model has full symmetry \( S_3 \).

Example 2:

We recognize this as the cyclic group \( C_3 \) with generator \( a_2 \) satisfying \( a_3 = a_2^2 \) and identity element \( a_1 = a_2^3 \). Taking the regular representation produces the permutation matrices

\[
A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.
\]

In this case, \( L_1 = -I + A_1 = 0 \), so we obtain the two-dimensional Lie-Markov model:

\[
Q = \alpha L_2 + \beta L_3 = \begin{pmatrix} -\alpha - \beta & \beta & \alpha \\ \alpha & -\alpha - \beta & \beta \\ \beta & \alpha & -\alpha - \beta \end{pmatrix}.
\]

We can also interpret this as the group-based model obtained from the group \( \mathbb{Z}_3 \cong C_3 \) with matrix entries satisfying \( q_{ij} = f(i - j) \) where the row and column indices \( i, j \in \mathbb{Z}_3 \), \( f(1) = \beta \) and \( f(2) = \alpha \).

This model has full symmetry \( S_3 \).

Example 3: Illustrating another example of a semigroup-based model with an absorbing state, consider the semigroup:

We take the regular representation produces the matrices:

\[
A_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}.
\]
and hence the semigroup-based model:

\[ Q = \alpha L_1 + \beta L_2 + \gamma L_3 = \begin{pmatrix} -\beta - \gamma & \alpha & 0 \\ \beta & -\alpha - \gamma & 0 \\ \gamma & \gamma & 0 \end{pmatrix}, \]

for which the third state is an absorbing state. The resulting Lie algebra has commutators:

\[ [L_1, L_2] = L_1 - L_2, \quad [L_1, L_3] = 0, \quad [L_2, L_3] = 0. \]

Due to the distinguished third state, this model has symmetry group \( \{ e, (12) \} < S_3. \)

5 Discussion

We did not attempt to derive semigroup-based models for any more than the \( k = 4 \) state case. One obstruction to going further is the combinatorial explosion in the number non-isomorphic semigroups of degree \( k \) (OEIS sequence A027851):

\[ 1, 5, 24, 188, 1915, 28634, 1627672, 3684030417, \ldots. \]

Clearly, it becomes computationally infeasible to systematically explore all possible cases as we have done in this paper. One possible method to proceed further, is to reject all those semigroups that do not satisfy a reasonable set of permutation symmetries. For a semigroup \( S \), the appropriate definition of a symmetry corresponds to the existence of an automorphism (self-isomorphism) \( \varphi : S \to S \). It is clear that the reasonable semigroup-based models we have derived in this paper are based on semigroups with non-trivial automorphism groups. However, the precise connection to reducibility of the so-derived Lie-Markov model is not clear.

Another reason we did not extend our constructions further beyond the \( k = 4 \) case is due to our particular interest in phylogenetics and DNA substitution models. Thus, although the primary focus in this paper was mathematical, it is worthwhile to ponder whether the new four-state model we obtained is useful for the analysis of real data sets.

Finally, as alluded to at the end of the introduction, it is likely at least some of the mathematical appealing features of the equal-input model [discussed in detail in Steel (2016, Chap 7)] are primarily due to the underlying semigroup identified in this work. As confirmed explicitly for the \( C_2 \times B_2 \) model in Sect. 4.2, we conjecture that similar mathematical tractability extends to any semigroup-based model.

We leave all of the above to future work.

Acknowledgements This research was supported by Australian Research Council (ARC) Discovery Grant DP150100088. We would like to thank Des FitzGerald for helpful comments on an early draft, and the anonymous reviewers for their careful reading and suggestions that lead to a substantially improved manuscript.
References

Casanellas M, Sullivant S (2005) The strand symmetric model. In: Algebraic statistics for computational biology. Cambridge University Press, New York, pp 305–321
Draisma J, Kuttler J (2009) On the ideals of equivariant tree models. Math Ann 344(3):619–644
Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6):368–376
Fernández-Sánchez J, Sumner JG, Jarvis PD, Woodhams MD (2015) Lie markov models with purine/pyrimidine symmetry. J Math Biol 70(4):855–891
Forsythe GE (1955) SWAC computes 126 distinct semigroups of order 4. Proc Am Math Soc 6(3):443–447
Hall BC (2015) Lie groups, Lie algebras, and representations: an elementary introduction. Springer, Berlin
Hasegawa M, Kishino H, Yano T (1985) Dating of human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174
Hendy MD, Penny D, Steel MA (1994) A discrete Fourier analysis for evolutionary trees. Proc Natl Acad Sci 91(8):3339–3343
Jarvis P, Sumner J (2012) Markov invariants for phylogenetic rate matrices derived from embedded sub-models. IEEE/ACM Trans Comput Biol Bioinform 9(3):828–836
Jarvis PD, Sumner JG (2016) Matrix group structure and markov invariants in the strand symmetric phylogenetic substitution model. J Math Biol 73(2):259–282. https://doi.org/10.1007/s00285-015-0951-7
Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111–120
Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci 78(1):454–458
Kingman JFC (1962) The imbedding problem for finite Markov chains. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 1(1):14–24
Kolmogorov A (1936) Zur Theorie der Markoffschen Ketten. Math Ann 112:155–160
Semple C, Steel MA (2003) Phylogenetics, vol 24. Oxford University Press, Oxford
Steel M (2016) Phylogeny: discrete and random processes in evolution. CBMS-NSF regional conference series on mathematics, vol 89. SIAM, 293 pp
Stillwell J (2008) Naive lie theory. Springer, Berlin
Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12(2):204–228
Sumner JG (2013) Lie geometry of 2 × 2 markov matrices. J Theor Biol 327(21):88–90
Sumner JG (2017) Multiplicatively closed markov models must form Lie algebras. ANZIAM J 59(2):240–246
Sumner JG, Fernández-Sánchez J, Jarvis PD (2012a) Lie Markov models. J Theor Biol 298:16–31
Sumner JG, Jarvis PD, Fernández-Sánchez J, Kaine BT, Woodhams MD, Holland BR (2012b) Is the general time-reversible model bad for molecular phylogenetics? Syst Biol 61(6):1069–1074
Székely LA, Steel MA, Erdős PL (1993) Fourier calculus on evolutionary trees. Adv Appl Math 14(2):200–216
Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci (Am Soc) 17:57–86
Woodhams MD, Fernández-Sánchez J, Sumner JG (2015) A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. Syst Biol 64(4):638–650
Yap VB, Pachter L (2004) Identification of evolutionary hotspots in the rodent genomes. Genome Res 14(4):574–579