Ardiyansyah, Muhammad; Kosta, Dimitra; Kubjas, Kaie

The model-specific Markov embedding problem for symmetric group-based models

Published in:
Journal of Mathematical Biology

DOI:
10.1007/s00285-021-01656-5

Published: 09/09/2021

Document Version
Publisher's PDF, also known as Version of record

Journal of Mathematical Biology (2021) 83:33
https://doi.org/10.1007/s00285-021-01656-5

Mathematical Biology

The model-specific Markov embedding problem for symmetric group-based models

Muhammad Ardiyansyah[1] · Dimitra Kosta[2] · Kaie Kubjas[1]

Received: 15 June 2020 / Revised: 4 August 2021 / Accepted: 17 August 2021 /
Published online: 9 September 2021

Abstract
We study model embeddability, which is a variation of the famous embedding problem in probability theory, when apart from the requirement that the Markov matrix is the matrix exponential of a rate matrix, we additionally ask that the rate matrix follows the model structure. We provide a characterisation of model embeddable Markov matrices corresponding to symmetric group-based phylogenetic models. In particular, we provide necessary and sufficient conditions in terms of the eigenvalues of symmetric group-based matrices. To showcase our main result on model embeddability, we provide an application to hachimoji models, which are eight-state models for synthetic DNA. Moreover, our main result on model embeddability enables us to compute the volume of the set of model embeddable Markov matrices relative to the volume of other relevant sets of Markov matrices within the model.

Keywords Embedding problem · Evolutionary models · Group-based models · Markov matrix · Markov generator

Mathematics Subject Classification 60J27 · 15A16 · 92D15 · 62R01

✉ Kaie Kubjas
kaie.kubjas@aalto.fi

Muhammad Ardiyansyah
muhammad.ardiyansyah@aalto.fi

Dimitra Kosta
D.Kosta@ed.ac.uk

[1] Department of Mathematics and Systems Analysis, Aalto University, Espoo, Finland

[2] School of Mathematics, University of Edinburgh, Edinburgh, UK

1 Introduction

The embedding problem for stochastic matrices, also known as Markov matrices, deals with the question of deciding whether a stochastic matrix M is the matrix exponential of a rate matrix Q. A rate matrix, also known as a Markov generator, has non-negative non-diagonal entries and row sums equal to zero. If a stochastic matrix satisfies such a property and can be expressed as a matrix exponential of a rate matrix, namely $M = e^{Qt}$, then M is called embeddable. Applications of the embeddability property vary from biology and nucleotide substitution models (Verbyla et al. 2013) to mathematical finance (Israel et al. 2001). For a first formulation of the embedding problem see Elfving (1937). An account of embeddable Markov matrices is provided in Davies (2010). The embedding problem for 2×2 matrices is due to Kendall and first published by Kingman (1962), for 3×3 matrices is fully settled in a series of papers Carette (1995), Chen and Chen (2011), Cuthbert (1973), Israel et al. (2001) and Johansen (1974), while for 4×4 stochastic matrices has been recently solved in Casanellas et al. (2020b). In general, when the size n of the stochastic matrix is greater than 4, the work Casanellas et al. (2020b) establishes a criterion for deciding whether a generic $n \times n$ Markov matrix with distinct eigenvalues is embeddable and proposes an algorithm that lists all its Markov generators. In the present paper we study a refinement of the classical embedding problem, called the model embedding problem for a class of $n \times n$ stochastic matrices coming from phylogenetic models.

Phylogenetics is the field that aims at reconstructing the history of evolution of species. A phylogenetic model is a mathematical model used to understand the evolutionary process given genetic data sets. The most popular phylogenetic models are nucleotide substitution models which use aligned DNA sequence data to study the molecular evolution of DNA. A comprehensive treatment of phylogenetic methods is given by Felsenstein, who is considered the initiator of statistical phylogenetics, in his seminal book Felsenstein (2003). Algebraic and geometric methods have been employed with great success in the study of phylogenetic models leading to an explosion of related research work and the establishment of the field of phylogenetic algebraic geometry, also known as algebraic phylogenetics; see Allman and Rhodes (2003), Baños et al. (2016), Casanellas and Fernández-Sánchez (2007), Cavender and Felsenstein (1987), Gross and Long (2018), Evans and Speed (1993), Lake (1987), Pachter and Sturmfels (2004) and Sturmfels and Sullivant (2005) for a non-exhaustive list of publications.

To build such a phylogenetic model, we first require a phylogenetic tree T, which is a directed acyclic graph comprising of vertices and edges representing the evolutionary relationships of a group of species. The vertices with valency 1 are called the leaves of the tree. The tree is considered rooted and the direction of evolution is from the root towards the leaves. On each vertex of the tree T, we associate a random variable with k possible states, where in phylogenetics k is often taken to be 2, for the binary states $\{0, 1\}$, or 4, to represent the four types of DNA nucleotides $\{\texttt{A, C, G, T}\}$. We also require a transition matrix (also known as a mutation matrix) $M^{(e)}$ corresponding to each edge e of the tree, where the entries of this $k \times k$ matrix $M^{(e)}$ represent the probabilities of transition between states. In a phylogenetic tree, the leaves correspond to extant species and so the random variables at the leaves are observed, while the

interior vertices correspond to possibly extinct species and so the random variables at the interior vertices are hidden.

In this paper we are focusing on symmetric group-based substitution models. Substitution models are a class of phylogenetic models which use a Markov process to describe the substitution of nucleotides over time in a given DNA sequence and for which the transition matrices along an edge e are stochastic matrices of the form $M^{(e)} = \exp\left(t_e Q^{(e)}\right)$. Group-based models are a special class of substitution models, in which the matrices $Q^{(e)}$ can be pairwise distinct, but they can all be simultaneously diagonalizable by a linear change of coordinates given by the discrete Fourier transform of an abelian group, also called a commutative group. For example, the Cavender–Farris–Neyman (CFN) model (Cavender 1978; Farris 1973; Neyman 1971), as well as the Jukes–Cantor (JC) (Jukes and Cantor 1969), the Kimura-2 parameter (K2P) (Kimura 1980) and the Kimura-3 parameter (K3P) (Kimura 1981) models for DNA are all group-based phylogenetic models. In Sturmfels and Sullivant (2005), it is established that through the discrete Fourier transform group-based models correspond to toric varieties, which are geometric objects with nice combinatorial properties. We are interested in symmetric group-based substitution models. Namely, apart from distinct and simultaneously diagonalizable, the transition matrices $Q^{(e)}$ are also symmetric square matrices. The symmetricity assumption guarantees that the eigenvalues of rate and transition matrices of a group-based model are real, a property that we use in the proof of our main theorem. Symmetric models are a subset of a special class of models called time-reversible models, where the Markov process appears identical when moving forward or backward in time. Our results apply to group-based models following an ergodic time-reversible Markov process, as in this case the rate matrices Q are symmetric according to Pachter and Sturmfels (2004, Lemma 17.2).

The classical embedding problem is concerned with deciding which square Markov matrices are embeddable, namely given a Markov matrix M whether there exists a rate matrix Q such that $M = \exp(Q)$. A variant of the embedding problem that asks for a reversible Markov generator for a stochastic matrix is studied in Jia (2016). When we impose the assumption that the rate matrix Q follows the corresponding model conditions, we arrive at a different refined notion of embeddability called model embeddability. The embeddability of circulant and equal-input stochastic matrices is studied in Baake and Sumner (2020). In the current paper, we focus on the $(\mathcal{G}, L)$-embeddability for $n \times n$ matrices corresponding to symmetric group-based substitution models. The $(\mathcal{G}, L)$-embeddability means that we require that the rate matrices Q preserve the symmetric group-based structure imposed by the abelian group $\mathcal{G}$ and the symmetric labelling L, which we define at the beginning of Sect. 2. Model embeddability for symmetric group-based models is relevant both for homogeneous and inhomogeneous time-continuous processes, as group-based models are Lie Markov models, and hence multiplicatively closed (Sumner et al. 2012; Verbyla et al. 2013). A study of the set of embeddable and model-embeddable matrices corresponding to the Jukes–Cantor, Kimura-2 and Kimura-3 DNA substitution models, which are all symmetric group-based models, is undertaken in Casanellas et al. (2020a) and Roca-Lacostena and Fernández-Sánchez (2018). In particular, a full characterisation of the set of embeddable 4×4 Kimura 2-parameter matrices is provided in Casanellas et al.

(2020a), which together with the results of Roca-Lacostena and Fernández-Sánchez (2018) fully solve the embedding problem for the Kimura 3-parameter model as well. Although model embeddability, which is a refined notion of embeddability imposed by the model structure, implies classical embeddability, the converse is generally not true (see also Roca-Lacostena and Fernández-Sánchez 2018, Example 3.1).

The main result of this paper is a characterization of $(\mathcal{G}, L)$-embeddability for any abelian group $\mathcal{G}$ equipped with a symmetric $\mathcal{G}$-labeling function L in Theorem 1. We provide necessary and sufficient conditions which the eigenvalues of the stochastic matrix of the model need to satisfy for the matrix to be $(\mathcal{G}, L)$-embeddable. To showcase our result, we first introduce three group-based models with the underlying group $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, based on the hachimoji DNA system introduced in Hoshika et al. (2019). Hachimoji, a Japanese word meaning "eight letters", is used to describe a synthetic analog of the nucleic acid DNA, where we have the four natural nucleobases {A,C,G,T} and furthermore an additional four synthetic nucleotides {P,Z,B,S}. We then apply Theorem 1 to characterise the model embeddability for the three hachimoji DNA models. The three models are called hachimoji 7-parameter, hachimoji 3-parameter and hachimoji 1-parameter models, which can be thought of as generalisations of the Kimura 3, Kimura 2 and Jukes–Cantor models respectively. Finally, the characterisation of model embeddability in terms of eigenvalues enables us to compute the volume of the $(\mathcal{G}, L)$-embeddable Markov matrices and compare this volume with volumes of other relevant sets of Markov matrices. For the general Jukes–Cantor model, which includes the hachimoji 1-parameter model, the volumes can be derived exactly; for the hachimoji 3-parameter model and for the hachimoji 7-parameter model symbolically and numerically.

The outline of the paper is the following. Section 2 gives a mathematical background covering notions such as the labeling functions, group-based models and the discrete Fourier transform. Section 3 introduces symmetric $\mathcal{G}$-compatible labelings which is a class of labeling functions with particularly nice properties. Section 4 presents the main result of this paper about the model embedding problem for symmetric group-based models equipped with a certain labeling function. Then in Sect. 5 we focus on the hachimoji DNA and provide exact characterisation of the model embeddability in terms of eigenvalues of the Markov matrix for the hachimoji 7-parameter, the hachimoji 3-parameter and the hachimoji 1-parameter models. Finally, Sect. 6 presents results on the volume of stochastic matrices that are $(\mathcal{G}, L)$-embeddable for the three hachimoji group-based models. The code for the computations in this paper is available at https://github.com/ardiyam1/Model-Embeddability-for-Symmetric-Group-Based-Models.

2 Preliminaries

In this section, we give background on group-based models and the discrete Fourier transform.

Definition 1 Let $\mathcal{G}$ be a finite additive abelian group and $\mathcal{L}$ a finite set. A *labeling function* is any function $L : \mathcal{G} \to \mathcal{L}$.

In the group-based model with underlying finite abelian group $\mathcal{G}$, states are in bijection with the elements of the group $\mathcal{G}$. Fundamental in the definition of a group-based model associated to a finite additive abelian group $\mathcal{G}$ and a labeling function L is that the rate of mutation from a state g to state h depends only on $L(h-g)$: That is, the entries of a rate matrix Q are

$$Q_{g,h} = \psi(h-g)$$

for a vector $\psi \in \mathbb{R}^{\mathcal{G}}$ satisfying $\sum_{g\in\mathcal{G}} \psi(g) = 0$, $\psi(g) \geq 0$ for all non-zero $g \in \mathcal{G}$ and $\psi(g) = \psi(h)$, whenever $L(g) = L(h)$. We say that such Q is a $(\mathcal{G}, L)$-rate matrix. In this paper, the rate matrices in group-based models are assumed to be symmetric, i.e., $\psi(-g) = \psi(g)$ for every $g \in \mathcal{G}$. Since the matrix exponential of a symmetric matrix is again symmetric, then the entries of a transition matrix $P = \exp(Q)$ are

$$P_{g,h} = f(h-g)$$

for a vector $f \in \mathbb{R}^{\mathcal{G}}$ satisfying $\sum f(g) = 1$, $f(g) \geq 0$ for all $g \in \mathcal{G}$ and $f(g) = f(-g)$ for all $g \in \mathcal{G}$. As we see in Example 2, in general it is not true that $f(g) = f(h)$ whenever $L(g) = L(h)$. In Sect. 3, we introduce $\mathcal{G}$-compatible labeling functions which guarantee this property and then we say that P is a $(\mathcal{G}, L)$-Markov matrix.

Example 1 Let $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$ and $\mathcal{L} = \{0, 1, 2, 3\}$. We identify nucleotides with the group elements of $\mathbb{Z}_2 \times \mathbb{Z}_2$ as $\mathtt{A} = (0,0), \mathtt{T} = (0,1), \mathtt{C} = (1,0)$ and $\mathtt{G} = (1,1)$. The Kimura 3-parameter, the Kimura 2-parameter and the Jukes–Cantor models correspond to the following labeling functions

$$\begin{aligned}
&L((0,0)) = 0, \quad L((0,1)) = 1, \quad L((1,0)) = 2, \quad L((1,1)) = 3,\\
&L((0,0)) = 0, \quad L((0,1)) = L((1,0)) = 1, \quad L((1,1)) = 2,\\
&L((0,0)) = 0, \quad L((0,1)) = L((1,0)) = L((1,1)) = 1,
\end{aligned}$$

respectively. The Kimura 3-parameter rate and transition matrices have the form

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}. \tag{2.1}$$

In the case of the Kimura 2-parameter model additionally $b = c$, after choosing the ordering $\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}$ of nucleotide bases. This will be further explained in Example 3. The Jukes–Cantor model is the submodel when $b = c = d$.

Example 2 Let $\mathcal{G} = \mathbb{Z}_7$ and L be a labeling function such that $L(1) = L(2) = L(5) = L(6)$ and $L(3) = L(4)$. Consider the $(\mathcal{G}, L)$-rate matrix

$$Q = \begin{pmatrix} -1 & 0.125 & 0.125 & 0.25 & 0.25 & 0.125 & 0.125 \\ 0.125 & -1 & 0.125 & 0.125 & 0.25 & 0.25 & 0.125 \\ 0.125 & 0.125 & -1 & 0.125 & 0.125 & 0.25 & 0.25 \\ 0.25 & 0.125 & 0.125 & -1 & 0.125 & 0.125 & 0.25 \\ 0.25 & 0.25 & 0.125 & 0.125 & -1 & 0.125 & 0.125 \\ 0.125 & 0.25 & 0.25 & 0.125 & 0.125 & -1 & 0.125 \\ 0.125 & 0.125 & 0.25 & 0.25 & 0.125 & 0.125 & -1 \end{pmatrix}.$$

In this rate matrix, $\psi(1) = \psi(2) = \psi(5) = \psi(6) = 0.125$ and $\psi(3) = \psi(4) = 0.25$. By direct computation, we get

$$P = e^{Q} = \begin{pmatrix} 0.41305 & 0.0858551 & 0.0834148 & 0.124205 & 0.124205 & 0.0834148 & 0.0858551 \\ 0.0858551 & 0.41305 & 0.0858551 & 0.0834148 & 0.124205 & 0.124205 & 0.0834148 \\ 0.0834148 & 0.0858551 & 0.41305 & 0.0858551 & 0.0834148 & 0.124205 & 0.124205 \\ 0.124205 & 0.0834148 & 0.0858551 & 0.41305 & 0.0858551 & 0.0834148 & 0.124205 \\ 0.124205 & 0.124205 & 0.0834148 & 0.0858551 & 0.41305 & 0.0858551 & 0.0834148 \\ 0.0834148 & 0.124205 & 0.124205 & 0.0834148 & 0.0858551 & 0.41305 & 0.0858551 \\ 0.0858551 & 0.0834148 & 0.124205 & 0.124205 & 0.0834148 & 0.0858551 & 0.41305 \end{pmatrix}.$$

The matrix P is not a $(\mathcal{G}, L)$-Markov matrix, since $0.0858551 = f(1) = f(6) \neq f(2) = f(5) = 0.0834148$. The entries of P induce a labeling function L' such that $L'(1) = L'(6) \neq L'(2) = L'(5)$ and $L'(3) = L'(4)$. In this case, the matrix P is a $(\mathcal{G}, L')$-Markov matrix.

Example 2 shows that the matrix exponential does not necessarily preserve the labeling function associated to a rate matrix. Conversely, Example 3.1 of Roca-Lacostena and Fernández-Sánchez (2018) suggests that a Kimura 3-parameter Markov matrix can be embeddable, despite the fact that it does not have any Markov generator satisfying Kimura 3-parameter model constraints.

Let $\mathbb{C}^*$ denote the multiplicative group of complex numbers without zero. A group homomorphism from $\mathcal{G}$ to $\mathbb{C}^*$ is called a *character* of $\mathcal{G}$. The characters of $\mathcal{G}$ form a group under multiplication, called the *character group* of $\mathcal{G}$ and denoted by $\widehat{\mathcal{G}}$. Here the product of two characters χ_1, χ_2 of the group $\mathcal{G}$ is defined by $(\chi_1\chi_2)(g) = \chi_1(g)\chi_2(g)$ for every $g \in \mathcal{G}$. The character group $\widehat{\mathcal{G}}$ is isomorphic to $\mathcal{G}$. Given a group isomorphism between $\mathcal{G}$ and $\widehat{\mathcal{G}}$, we will denote by $\widehat{g} \in \widehat{\mathcal{G}}$ the image of $g \in \mathcal{G}$. For a finite group $\mathcal{G}$, the values of characters are roots of unity.

Lemma 1 (Pachter and Sturmfels 2005, Lemma 17.1) *Let $g, h \in \mathcal{G}$ and $k \in \mathbb{Z}$. Then $\widehat{g}(-h) = \overline{\widehat{g}(h)}$ and $\widehat{kg}(h) = \widehat{g}(kh)$, where $\overline{a}$ denotes the complex conjugate of $a \in \mathbb{C}$.*

Given a function $a : \mathcal{G} \to \mathbb{C}$, its *discrete Fourier transform* is a function $\check{a} : \mathcal{G} \to \mathbb{C}$ defined by

$$\check{a}(g) := \sum_{h \in \mathcal{G}} \widehat{g}(h) a(h).$$

Lemma 2 (Matsen 2008, Section 2) *For any real-valued function $a : \mathcal{G} \to \mathbb{C}$, the identity $\check{a}(-g) = \overline{\check{a}(g)}$ holds for all $g \in \mathcal{G}$. Moreover, if $a(-g) = a(g)$ for all $g \in \mathcal{G}$, then $\check{a}(-g) = \check{a}(g)$ for all $g \in \mathcal{G}$ and $\check{a}$ is a real-valued function.*

In the proof of Theorem 1, we will use that $\check{\psi}$ and $\check{f}$ are real-valued. For this reason, in this paper we consider only group-based models that are equipped with *symmetric* labeling functions, i.e. $L(g) = L(-g)$ for all $g \in \mathcal{G}$. In other words, a symmetric group-based model assumes that the transition matrices are real symmetric matrices.

The discrete Fourier transform is a linear endomorphism on $\mathbb{C}^{\mathcal{G}}$. We will denote its matrix by K. In particular, the entries of K are $K_{g,h} = \widehat{g}(h)$ for $g, h \in \mathcal{G}$. The matrix K is symmetric for any finite abelian group (Luong 2009, Section 3.2). The inverse of the discrete Fourier transformation matrix is $K^{-1} = \frac{1}{|\mathcal{G}|}K^*$, where K^* denotes the adjoint of K (Luong 2009, Corollary 3.2.2).

The following lemma describes the relation between functionals f and ψ in the case $f(-g) = f(g)$ and $\psi(-g) = \psi(g)$ for all $g \in \mathcal{G}$.

Lemma 3 (Matsen 2008, Lemma 2.2) *Let Q be determined by $\psi \in \mathbb{R}^{\mathcal{G}}$ and P be determined by $f \in \mathbb{R}^{\mathcal{G}}$ as described earlier in this section such that $P = e^Q$. Furthermore, assume that $\psi(g) = \psi(-g)$ and $f(g) = f(-g)$ for all $g \in \mathcal{G}$. Then, $\check{f}(g) = e^{\check{\psi}(g)}$ for all $g \in \mathcal{G}$.*

Lemma 4 *Let Q be determined by $\psi \in \mathbb{R}^{\mathcal{G}}$ and P be determined by $f \in \mathbb{R}^{\mathcal{G}}$ as described earlier in this section. Furthermore, assume that $\psi(g) = \psi(-g)$ and $f(g) = f(-g)$ for all $g \in \mathcal{G}$. Let K_g denote the column of the discrete Fourier transform matrix labeled by g. The eigenpairs of Q (resp. P) are $(\check{\psi}(g), K_g)$ (resp. $(\check{f}(g), K_g)$) for $g \in \mathcal{G}$.*

Proof This result is stated in the proof of Matsen (2008, Lemma 2.2).

In particular, in the case of a Markov matrix, the column vector of ones is an eigenvector with eigenvalue one. In the case of a rate matrix, the column vector of ones is an eigenvector with eigenvalue zero. A direct consequence of Lemma 4 is that Q and P are diagonalizable by K, i.e. $Q = KD_1K^{-1}$ and $P = KD_2K^{-1}$ where D_1 and D_2 are diagonal matrices with diagonals given by the vectors $\check{\psi}$ and $\check{f}$ of $\mathbb{R}^{\mathcal{G}}$ respectively.

3 $\mathcal{G}$-compatible labeling functions

In this section, we introduce a class of labeling functions with the property that the symmetries of the probability vector are preserved under the discrete Fourier transformation. This property is required for any result that is proven using the discrete Fourier transform. Notably, labeling functions for all common group-based models (CFN, K3P, K2P, and JC models) are $\mathcal{G}$-compatible.

Definition 2 Let $\mathcal{G}$ be a finite additive abelian group, $\mathcal{L}$ a set and $L : \mathcal{G} \to \mathcal{L}$ a labeling function. Let K be the discrete Fourier transformation matrix for $\mathcal{G}$ and x_L be the column vector of length $|\mathcal{G}|$ whose g-th component is the indeterminate $x_{L(g)}$. We say that L is a $\mathcal{G}$-compatible labeling function if for every $g, h \in \mathcal{G}$ with $L(g) = L(h)$, we have that $K_{g,:} \cdot x_L = K_{h,:} \cdot x_L$ and $(K^{-1})_{g,:} \cdot x_L = (K^{-1})_{h,:} \cdot x_L$. Here $M_{a,:}$ denotes the row of M indexed by group element a.

A labeling function that maps every group element to a different label is trivially $\mathcal{G}$-compatible.

Remark 1 In the definition of a $\mathcal{G}$-compatible labeling, we require that the matrices K and K^{-1} preserve the symmetries of the vector of labels x_L. For symmetric group-based models, it is enough to require that only K or K^{-1} preserves the symmetries of the vector of labels x_L. Recall that

$$K^{-1} \cdot x_L = \frac{1}{|\mathcal{G}|} \cdot K^* \cdot x_L = \frac{1}{|\mathcal{G}|} \cdot \overline{K} \cdot x_L.$$

The property $\widehat{g}(-h) = \overline{\widehat{g}(h)}$ implies $K_{g,-h} = \overline{K_{g,h}}$ and $\overline{K_{g,-h}} = K_{g,h}$. If $-h = h$, this means $\overline{K_{g,h}} = K_{g,h}$ for all $g \in \mathcal{G}$. If $-h \neq h$, then taking into account that $x_{L(-h)} = x_{L(h)}$ gives $\overline{K_{g,h}} \cdot x_{L(h)} + \overline{K_{g,-h}} \cdot x_{L(-h)} = K_{g,h} \cdot x_{L(h)} + K_{g,-h} \cdot x_{L(-h)}$. Hence $K^{-1} \cdot x_L = 1/|\mathcal{G}| \cdot K \cdot x_L$.

Remark 2 If a labeling function L is symmetric $\mathcal{G}$-compatible and Q is a $(\mathcal{G}, L)$-rate matrix, then a Markov matrix $P = e^Q$ is a $(\mathcal{G}, L)$-Markov matrix, i.e. $P_{g,h} = f(h-g)$ for a vector $f \in \mathbb{R}^G$ and $f(g) = f(h)$ whenever $L(g) = L(h)$.

Example 3 Let $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$. The discrete Fourier transformation matrix for $\mathcal{G}$ is

$$K = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

To show $\mathcal{G}$-compatibility for the three labeling functions from Example 1, it is enough to check that K preserves the symmetries of x_L. The labeling function of the Jukes–Cantor model is $\mathcal{G}$-compatible, since

$$K \cdot \begin{pmatrix} x_0 \\ x_1 \\ x_1 \\ x_1 \end{pmatrix} = \begin{pmatrix} x_0 + 3x_1 \\ x_0 - x_1 \\ x_0 - x_1 \\ x_0 - x_1 \end{pmatrix}.$$

The labeling function of the Kimura 2-parameter model is $\mathcal{G}$-compatible, since

$$K \cdot \begin{pmatrix} x_0 \\ x_1 \\ x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_0 + 2x_1 + x_2 \\ x_0 - x_2 \\ x_0 - x_2 \\ x_0 - 2x_1 + x_2 \end{pmatrix}.$$

In the literature, usually $L((1,0)) = L((1,1))$ in the Kimura 2-parameter model. However, here we assume that $L((1,0)) = L((0,1))$ which is simply due to the fact that we consider the identification $\mathtt{A} = (0,0)$, $\mathtt{T} = (0,1)$, $\mathtt{C} = (1,0)$ and $\mathtt{G} = (1,1)$. To be more precise, nucleotide bases fall into two categories depending on

the molecular mechanisms of the base; purines (A or G) and pyrimidines (C or T). A transition occurs when a purine is substituted by a purine, or a pyrimidine by a pyrimidine. A change from a purine to a pyrimidine, or vice versa, is a transversion. The Kimura 2-parameter model of sequence evolution distinguishes between transitions and transversions to account for the biological fact that transitions occur at higher rate than transversions (Kimura 1980, 1981). The rate and transition matrix of the Kimura 2-parameter model have the form

$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{T} & \mathtt{C} & \mathtt{G} \end{array} \\ \begin{pmatrix} a & b & b & d \\ b & a & d & b \\ b & d & a & b \\ d & b & b & a \end{pmatrix} \end{array}.$$

The reason for choosing this identification and ordering is that we can use the discrete Fourier transform matrix in a format, which better demonstrates that it is the Kronecker product of discrete Fourier transformation matrices for $\mathbb{Z}_2$. The labeling function of the Kimura 3-parameter is $\mathcal{G}$-compatible, because each group element maps to a different label.

Sturmfels and Sullivant consider a different class of labeling functions, called friendly labelings (Sturmfels and Sullivant 2005). Group-based models with friendly labeling functions are equivalent to $\mathcal{G}$-models defined by Michałek (2011, Remark 5.2). $\mathcal{G}$-models are constructed using an arbitrary group $\mathcal{G}$ that has a normal, abelian subgroup $\mathcal{H}$ which acts transitively and freely on the finite set of states. The importance of $\mathcal{G}$-models is that they are toric. We explore connections between friendly labelings and $\mathcal{G}$-compatible labelings in "Appendix A". We conjecture that every symmetric $\mathcal{G}$-compatible labeling is a friendly labeling, but not vice versa.

The following lemma provides a necessary condition for $\mathcal{G}$-compatible labeling functions.

Lemma 5 *Let $\mathcal{G}$ be a finite additive abelian group, $\mathcal{L}$ a set and $L : \mathcal{G} \to \mathcal{L}$ a labeling function. If L is $\mathcal{G}$-compatible, then $L(0) \neq L(g)$ for any $g \neq 0$.*

Proof Let K be the discrete Fourier transformation matrix for $\mathcal{G}$. The entries of K are $\widehat{g}(h)$ for $g, h \in \mathcal{G}$, which are roots of unity. The row $K_{0,:}$ consists of ones. On the other hand, no other row of K consists of ones only, as this would contradict the uniqueness of the identity element in the character group. In particular, every other row of K contains at least one element whose real part is strictly less than one. Thus it is impossible that $K_{0,:} \cdot x_L = K_{g,:} \cdot x_L$ for $g \neq 0$.

Table 1 summarizes up to isomorphism all possible symmetric $\mathcal{G}$-compatible labeling functions for additive abelian groups of order up to eight. In the table, two group elements receive the same label if they belong to the same subset in a partition of $\mathcal{G}$.

We saw in Example 3 that the labeling function of the Jukes–Cantor model that assigns the same label to each nonzero element of the group $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$ is a $\mathcal{G}$-compatible labeling. This example can be generalized to other groups.

Table 1 Symmetric $\mathcal{G}$-compatible labelings for abelian groups of order $n \leq 8$ up to isomorphism

n	Group	Symmetric $\mathcal{G}$-compatible labelings
2	$\mathbb{Z}_2$	{{0},{1}}
3	$\mathbb{Z}_3$	{{0},{1,2}}
4	$\mathbb{Z}_4$	{{0},{1,2,3}},{{0},{1,3},{2}}
4	$\mathbb{Z}_2 \times \mathbb{Z}_2$	{{(0,0)},{(0,1),(1,0),(1,1)}},{{(0,0)},{(0,1),(1,0)},{(1,1)}}, {{(0,0)},{(0,1)},{(1,0)},{(1,1)}}
5	$\mathbb{Z}_5$	{{0},{1,2,3,4}}, {{0},{1,4},{2,3}}
6	$\mathbb{Z}_2 \times \mathbb{Z}_3$	{{(0,0)},{(0,1),(0,2),(1,0),(1,1),(1,2)}}, {{(0,0)},{(0,1),(0,2)},{(1,0)},{(1,1),(1,2)}}
7	$\mathbb{Z}_7$	{{0},{1,2,3,4,5,6}},{{0},{1,6},{2,5},{3,4}}
8	$\mathbb{Z}_8$	{{0},{1,2,3,4,5,6,7}},{{0},{1,3,5,7},{2,6},{4}}, {{0},{1,7},{2,6},{3,5},{4}}
8	$\mathbb{Z}_2 \times \mathbb{Z}_4$	{{(0,0)},{(0,1),(0,2),(0,3),(1,0),(1,1),(1,2),(1,3)}}, {{(0,0)},{(0,1),(0,3),(1,1),(1,3)},{(0,2)},{(1,0),(1,2)}}, {{(0,0)},{(0,1),(0,2),(0,3)},{(1,0)},{(1,1),(1,2),(1,3)}}, {{(0,0)},{(0,1),(0,3),(1,0)},{(0,2),(1,1),(1,3)},{(1,2)}}, {{(0,0)},{(0,1),(0,3)},{(0,2)},{(1,0)},{(1,1),(1,3)},{(1,2)}}
8	$\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$	{{(0,0,0)},{(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)}}, {{(0,0,0)},{(0,0,1)},{0,1,0),(1,0,0),(1,1,0)},{(0,1,1),(1,0,1),(1,1,1)}}, {{(0,0,0)},{(0,0,1),(1,0,0),(1,0,1)}, {(0,1,0)},{(0,1,1),(1,1,0),(1,1,1)}}, {{(0,0,0)},{(0,0,0),(0,1,0),(0,1,1)},{(1,0,0)},{(1,0,1),(1,1,0),(1,1,1)}}, {{(0,0,0)},{(0,0,1),(0,1,0),(1,0,1),(1,1,0)},{(0,1,1)},{(1,0,0),(1,1,1)}}, {{(0,0,0)},{(0,0,1),(0,1,0),(1,0,0)},{(0,1,1),(1,0,1),(1,1,0)},{(1,1,1)}}, {{(0,0,0)},{(0,0,1),(1,1,1)},{(0,1,0),(0,1,1),(1,0,0),(1,0,1)},{(1,1,0)}}, {{(0,0,0)},{(0,0,1),(0,1,1),(1,0,0),(1,1,0)},{(0,1,0),(1,1,1)},{(1,0,1)}}, {{(0,0,0)},{(0,0,1)},{(0,1,0),(1,0,0)},{(0,1,1),(1,0,1)},{(1,1,0)},{(1,1,1)}}, {{(0,0,0)},{(0,0,1),(0,1,0)},{(0,1,1)},{(1,0,0)},{(1,0,1),(1,1,0)},{(1,1,1}}, {{(0,0,0)},{(0,0,1),(1,0,0)},{(0,1,0)},{(0,1,1),(1,1,0)},{(1,0,1)},{(1,1,1}}, {{(0,0,0)},{(0,0,1)},{(0,1,0)},{(0,1,1)},{(1,0,0)},{(1,0,1)},{(1,1,0)},{(1,1,1)}}

Lemma 6 *Let $\mathcal{G}$ be a finite abelian group. Let $L : \mathcal{G} \to \{0, 1\}$ be a labeling function such that and $L(0) = 0$ and $L(g) = 1$ for $g \neq 0$. Then the labeling function L is symmetric $\mathcal{G}$-compatible.*

Proof Clearly the labeling function L is symmetric. Let K be the discrete Fourier transformation matrix for $\mathcal{G}$. By Luong (2009, Corollary 3.2.1), we have

$$\sum_{h \in \mathcal{G}} K_{g,h} = \begin{cases} n, & g = 0 \\ 0, & g \neq 0 \end{cases}.$$

Hence

$$K_{g,:} \cdot x_L = \begin{cases} x_0 + (n-1)x_1, & g = 0 \\ x_0 - x_1, & g \neq 0 \end{cases}.$$

Hence L is a $\mathcal{G}$-compatible labeling function.

We call the model in Lemma 6 the *general Jukes–Cantor model*. We finish this section with giving another class of labeling functions that are $\mathcal{G}$-compatible for every finite abelian group $\mathcal{G}$.

Lemma 7 *Let $\mathcal{G}$ be a finite additive abelian group, $\mathcal{L}$ a finite set and $L : \mathcal{G} \to \mathcal{L}$ a labeling function such that for any two distinct elements $g, h \in \mathcal{G}$, $L(g) = L(h)$ if and only if $g = -h$. Then L is $\mathcal{G}$-compatible.*

Proof By Lemma 1, the identity $\widehat{-g}(h) = \widehat{g}(-h)$ holds for all $g, h \in \mathcal{G}$. Then

$$\begin{aligned} K_{-g,:} \cdot x_L &= \sum_{h\in\mathcal{G}} \widehat{-g}(h) x_{L(h)} = \sum_{h\in\mathcal{G}} \widehat{g}(-h) x_{L(h)} \\ &= \sum_{h\in\mathcal{G}} \widehat{g}(h) x_{L(-h)} = \sum_{h\in\mathcal{G}} \widehat{g}(h) x_{L(h)} = K_{g,:} \cdot x_L. \end{aligned}$$

Thus, the labeling function L is $\mathcal{G}$-compatible as x_L is the column vector of indeterminates $x_{L(g)}$.

The converse of Lemma 7 is not true in general. Two examples are given by the labeling functions for the Kimura 2-parameter and the Jukes–Cantor model.

4 Model embeddability

The following theorem is the main result of this paper. It characterizes $(\mathcal{G}, L)$-embeddable transition matrices in terms of their eigenvalues.

Theorem 1 *Fix a finite abelian group $\mathcal{G}$, a finite set $\mathcal{L}$, and a symmetric $\mathcal{G}$-compatible labeling function $L : \mathcal{G} \to \mathcal{L}$. Let P be a $(\mathcal{G}, L)$-Markov matrix. Then P is $(\mathcal{G}, L)$-embeddable if and only if the vector $\lambda \in \mathbb{R}^{\mathcal{G}}$ of eigenvalues of P is in the set*

$$\begin{aligned} \{\lambda \in \mathbb{R}^{\mathcal{G}} : &\lambda_0 = 1, \prod_{h\in\mathcal{G}} \lambda_h^{Re((K)_{g,h})} \geq 1 \text{ for all nonzero } g \in \mathcal{G}, \\ &\lambda_g > 0 \text{ for all } g \in \mathcal{G}, \text{ and } \lambda_g = \lambda_h \text{ whenever } L(g) = L(h)\}. \end{aligned}$$

Proof We start by summarizing the idea of the proof. We consider the set $\Psi_{\mathcal{G},L}$ that consists of vectors ψ that determine $(\mathcal{G}, L)$-rate matrices. Our goal is to characterize the set $\check{F}_{\mathcal{G},L}$ of eigenspectra of Markov matrices that are matrix exponentials of $(\mathcal{G}, L)$-rate matrices determined by vectors ψ in $\Psi_{\mathcal{G},L}$. The first step is to consider the discrete Fourier transform of the set $\Psi_{\mathcal{G},L}$, which we denote by $\check{\Psi}_{\mathcal{G},L}$. By Lemma 4, this set is the set of eigenvalues of the $(\mathcal{G}, L)$-rate matrices. The second step is to consider the image of the set $\check{\Psi}_{\mathcal{G},L}$ under coordinatewise exponentiation. This set is precisely $\check{F}_{\mathcal{G},L}$, because $(\mathcal{G}, L)$-rate matrices are diagonalizable by the discrete Fourier transform matrix K by the discussion after Lemma 4 and thus if a $(\mathcal{G}, L)$-rate matrix Q is determined by $\psi \in \mathbb{R}^{\mathcal{G}}$ then

$$P = e^Q = K \cdot e^{\mathrm{diag}(\check{\psi})} \cdot K^{-1} = K \cdot \mathrm{diag}(e^{\check{\psi}}) \cdot K^{-1},$$

where $\check{\psi}$ is the vector of eigenvalues of Q and $e^{\check{\psi}}$ is the vector of eigenvalues of P.

More specifically, let

$$\Psi_{\mathcal{G},L} = \{\psi \in \mathbb{R}^{\mathcal{G}} : \sum_{g\in\mathcal{G}} \psi(g) = 0,\ \psi(g) \geq 0 \text{ for all nonzero } g \in \mathcal{G}, \text{ and } \psi(g) = \psi(h) \text{ whenever } L(g) = L(h)\}.$$

The vectors in the set $\Psi_{\mathcal{G},L}$ are in one-to-one correspondence with $(\mathcal{G}, L)$-rate matrices. The image of $\Psi_{\mathcal{G},L}$ under the discrete Fourier transform is the set

$$\check{\Psi}_{\mathcal{G},L} = \{\check{\psi} \in \mathbb{R}^{\mathcal{G}} : \check{\psi}(0) = 0,\ (K^{-1}\check{\psi})(g) \geq 0 \text{ for all nonzero } g \in \mathcal{G}, \text{ and } \check{\psi}(g) = \check{\psi}(h) \text{ whenever } L(g) = L(h)\}.$$

By Lemma 4, this set is the set of eigenvalues of the $(\mathcal{G}, L)$-rate matrices.

The image of $\check{\Psi}_{\mathcal{G},L}$ under the coordinatewise exponentiation is the set of eigenvalues of the $(\mathcal{G}, L)$-Markov matrices, which we denote by $\check{F}_{\mathcal{G},L}$. We claim that $\check{F}_{\mathcal{G},L}$ is equal to the set

$$\{\check{f} \in \mathbb{R}^{\mathcal{G}} : \check{f}(0) = 1,\ \prod_{h\in\mathcal{G}} (\check{f}(h))^{(K^{-1})_{g,h}} \geq 1 \text{ for all nonzero } g \in \mathcal{G},\ \check{f}(g) > 0 \text{ for all } g \in \mathcal{G}, \text{ and } \check{f}(g) = \check{f}(h) \text{ whenever } L(g) = L(h)\}. \tag{4.1}$$

Indeed, let $\check{f} = \exp(\check{\psi})$. Then $\check{f} > 0$ because the image of the exponentiation map is positive. The inequality $a^T x \geq 0$ is equivalent to $\exp(a^T x) \geq 1$. Hence, the equation $\check{\psi}(0) = 0$ gives $\check{f}(0) = 1$ and the inequalities $(K^{-1}\check{\psi})(g) \geq 0$ give

$$\prod_{h\in\mathcal{G}} (\check{f}(h))^{(K^{-1})_{g,h}} = \prod_{h\in\mathcal{G}} (e^{(\check{\psi}(h))})^{(K^{-1})_{g,h}} = e^{\sum_{h\in\mathcal{G}} \check{\psi}(h)(K^{-1})_{g,h}} = e^{(K^{-1}\check{\psi})(g)} \geq 1 \tag{4.2}$$

for all nonzero $g \in \mathcal{G}$. Hence $\check{f}$ is in the set (4.1). Conversely, let $\check{f}$ be a vector in the set (4.1). Then under coordinatewise logarithm, $\log(\check{f}) \in \check{\Psi}_{\mathcal{G},L}$ and $\check{f} = \exp(\log(\check{f}))$. Hence $\check{f}$ is in the image of $\check{\Psi}_{\mathcal{G},L}$. Thus $\check{F}_{\mathcal{G},L}$ is equal to the set (4.1).

It is left to rewrite the inequalities (4.2) as in the statement of the theorem. We have

$$(K^{-1})_{g,-h} = \frac{1}{|\mathcal{G}|}\overline{K_{-h,g}} = \frac{1}{|\mathcal{G}|}\overline{\widehat{-h}(g)} = \frac{1}{|\mathcal{G}|}\overline{\widehat{h}(-g)} = \frac{1}{|\mathcal{G}|}\widehat{h}(g) = \frac{1}{|\mathcal{G}|}K_{h,g} = \overline{(K^{-1})_{g,h}}$$

for all $g, h \in \mathcal{G}$. Here we use Lemma 1 and the definition of the discrete Fourier transformation matrix. If $-h = h$, then $(K^{-1})_{g,h} = (K^{-1})_{g,-h} = \overline{(K^{-1})_{g,h}}$, and

hence $(K^{-1})_{g,h} = \mathrm{Re}((K^{-1})_{g,h})$. If $-h \neq h$, then $\check{f}(h) = \check{f}(-h)$ by Lemma 2. Hence

$$\begin{aligned}(\check{f}(h))^{(K^{-1})_{g,h}}(\check{f}(-h))^{(K^{-1})_{g,-h}} &= (\check{f}(h))^{(K^{-1})_{g,h}}(\check{f}(h))^{\overline{(K^{-1})_{g,h}}} \\ &= (\check{f}(h))^{2\mathrm{Re}((K^{-1})_{g,h})} \\ &= (\check{f}(h))^{\mathrm{Re}((K^{-1})_{g,h})}(\check{f}(-h))^{\mathrm{Re}((K^{-1})_{g,-h})}.\end{aligned} \tag{4.3}$$

We replace K^{-1} by $1/|\mathcal{G}| \cdot \overline{K}$ and take both sides of the resulting inequality to the power $|\mathcal{G}|$. Finally, making the substitution $\lambda_h = \check{f}(h)$ gives the desired characterization.

For $\mathcal{G}$ cyclic, Theorem 1 has been independently proven by Baake and Sumner in the context of circulant matrices (Baake and Sumner 2020, Theorem 5.7). Moreover, they show that every embeddable circulant matrix is circulant embeddable (Baake and Sumner 2020, Corollary 5.2).

It follows from Lemma 3 that if a $(\mathcal{G}, L)$-Markov matrix P is $(\mathcal{G}, L)$-embeddable, then there exists a unique $(\mathcal{G}, L)$-rate matrix Q such that $P = \exp(Q)$. Indeed, since Q and P have both real eigenvalues and the eigenvalues of P are exponentials of eigenvalues of Q, then the eigenvalues of Q are uniquely determined by the eigenvalues of P. Then the $(\mathcal{G}, L)$-rate matrix Q is the principal logarithm of P.

The inequalities $\lambda_g > 0$ in Theorem 1 imply $\det(P) = \prod \lambda_g > 0$. Hence the set of $(\mathcal{G}, L)$-embeddable matrices for a symmetric group-based model is a relatively closed subset of a connected component of the complement of $\det(P) = 0$. A relatively closed subset means here a set that can be written as the intersection of a closed subset of $\mathbb{R}^{\mathcal{G}\times\mathcal{G}}$ and the connected component of the complement of $\det(P) = 0$.

In the rest of the current section and in Sect. 5, we will discuss applications of Theorem 1. We will recover known results about $(\mathcal{G}, L)$-embeddability and as a novel application characterize embeddability for three group-based models of hachimoji DNA.

Example 4 The CFN model is the group-based model associated to the group $\mathbb{Z}_2$. The CFN Markov matrices have the form

$$P = \begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

The discrete Fourier transform matrix is

$$K = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The eigenvalues of P are $\lambda_0 = a + b = 1$ and $\lambda_1 = a - b$. By Theorem 1, the Markov matrix P is CFN embeddable if and only if $0 < \lambda_1 \leq 1$ or equivalently $0 < a - b \leq 1$. This is equivalent to P satisfying $\det(P) > 0$, or equivalently $\mathrm{tr}(P) > 1$. The result that a general 2×2 stochastic matrix is embeddable if and only

if $\det(P) > 0$ or $\mathrm{tr}(P) > 1$ goes back to Kingman (1962, Proposition 2). Hence P is CFN embeddable if and only if it is embeddable.

Example 5 Recall that the Kimura 3-parameter model is the group-based model associated to group $\mathcal{G} = \mathbb{Z}_2 \times \mathbb{Z}_2$ and a K3P Markov matrix P has the form (2.1). The eigenvalues of P are

$$\begin{aligned}&\lambda_{(0,0)} = a + b + c + d, \lambda_{(0,1)} = a - b + c - d, \lambda_{(1,0)}\\ &= a + b - c - d, \lambda_{(1,1)} = a - b - c + d.\end{aligned}$$

By Theorem 1, a Markov matrix P is K3P embeddable if and only if

$$\begin{aligned}&\lambda_{(0,0)} = 1, \lambda_{(0,1)} > 0, \lambda_{(1,0)} > 0, \lambda_{(1,1)} > 0,\\ &\lambda_{(0,1)} \geq \lambda_{(1,0)}\lambda_{(1,1)}, \lambda_{(1,0)} \geq \lambda_{(0,1)}\lambda_{(1,1)}, \lambda_{(1,1)} \geq \lambda_{(0,1)}\lambda_{(1,0)}.\end{aligned} \tag{4.4}$$

In the Kimura 2-parameter model $b = c$ and $\lambda_{(0,1)} = \lambda_{(1,0)}$. We get the conditions for the K2P embeddability by setting $\lambda_{(0,1)} = \lambda_{(1,0)}$ in (4.4). Hence a K2P Markov matrix is K2P embeddable if and only if

$$\lambda_{(0,0)} = 1, \lambda_{(0,1)} > 0, 1 \geq \lambda_{(1,1)} \geq \lambda_{(0,1)}^2.$$

In the Jukes–Cantor model $b = c = d$ and $\lambda_{(0,1)} = \lambda_{(1,0)} = \lambda_{(1,1)}$. A JC Markov matrix is JC embeddable if and only if

$$\lambda_{(0,0)} = 1, 1 \geq \lambda_{(0,1)} > 0.$$

The K3P embeddability of a K3P Markov matrix with no repeated eigenvalues is equivalent to the embeddability of the matrix. Similarly, the JC embeddability of a JC Markov matrix is equivalent to the embeddability of the matrix. The same is not true for K2P Markov matrices with exactly two coinciding eigenvalues. See Roca-Lacostena and Fernández-Sánchez (2018, Section 3) for similar computations and further discussion on the model embeddability of K3P, K2P, and JC Markov matrices.

Remark 3 By Kingman (1962, Corollary on page 18), the map from rate matrices to transition matrices is locally homeomorphic except possibly when the rate matrix has a pair of eigenvalues differing by a non-zero multiple of $2\pi i$. Since for symmetric group-based models rate matrices are real symmetric, then all their eigenvalues are real and hence the map from rate matrices to transition matrices is a homeomorphism. Therefore the boundaries of embeddable transition matrices of symmetric group-based models are images of the boundaries of the rate matrices. For general Markov model, the boundaries of embeddable transition matrices are characterized in Kingman (Kingman 1962, Propositions 5 and 6).

Corollary 1 *A $(\mathcal{G}, L)$-embeddable transition matrix lies on the boundary of the set of $(\mathcal{G}, L)$-embeddable transition matrices for a symmetric group-based model if and only if it satisfies at least one of the inequalities in Theorem 1 with equality.*

5 Hachimoji DNA

In this section, we suggest three group-based models for a genetic system with eight building blocks recently introduced by Hoshika et al. (2019), and then characterize model embeddability for the proposed group-based models. The genetic system is called *hachimoji DNA*. It has four synthetic nucleotides denoted S, B, Z, and P in addition to the standard nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T). Detailed descriptions of the four additional nucleotides are given in Hoshika et al. (2019). If in the standard 4-letter DNA, the purines are A and G and the pyrimidines are C and T, then in the hachimoji system, there are additionally purine analogs P and B, and pyrimidine analogs Z and S. The hydrogen bonds occur between the pairs A-T, C-G, S-B and Z-P.

This DNA genetic system with eight building blocks can reliably form matching base pairs and can be read and translated into RNA. It is mutable without damaging crystal structure which is required for molecular evolution. Hachimoji DNA has potential application in bar-coding, retrievable information storage, and self-assembling nanostructures.

The underlying group we suggest for the hachimoji DNA is $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, since when restricted to the standard 4-letter DNA it gives the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ that is the underlying group for the standard DNA models. We identify the nucleotides with the group elements of $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ as follows:

$$\begin{aligned} \mathtt{A} = (0,0,0), \mathtt{C} = (0,0,1), \mathtt{T} = (0,1,0), \mathtt{G} = (0,1,1), \\ \mathtt{P} = (1,0,0), \mathtt{Z} = (1,0,1), \mathtt{S} = (1,1,0), \mathtt{B} = (1,1,1). \end{aligned}$$

The discrete Fourier transformation matrix of the group $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ is

$$K = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}. \tag{5.1}$$

5.1 Hachimoji 7-parameter model

The first model we propose is the analogue of the Kimura 3-parameter model and we will call it the *hachimoji 7-parameter (H7P) model*. In the hachimoji 7-parameter model, each element of the group $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ maps to a distinct label. Thus the labeling function is trivially $(\mathcal{G}, L)$-compatible. The H7P rate and transition matrices have the form

$$\begin{pmatrix} a & b & c & d & e & f & g & h \\ b & a & d & c & f & e & h & g \\ c & d & a & b & g & h & e & f \\ d & c & b & a & h & g & f & e \\ e & f & g & h & a & b & c & d \\ f & e & h & g & b & a & d & c \\ g & h & e & f & c & d & a & b \\ h & g & f & e & d & c & b & a \end{pmatrix}. \tag{5.2}$$

The eigenvalues of a H7P Markov matrix are

$$\begin{aligned} &\left(1, \lambda_{(0,0,1)}, \lambda_{(0,1,0)}, \lambda_{(0,1,1)}, \lambda_{(1,0,0)}, \lambda_{(1,0,1)}, \lambda_{(1,1,0)}, \lambda_{(1,1,1)}\right)^T \\ &\quad = K \cdot \left(a, b, c, d, e, f, g, h\right)^T. \end{aligned}$$

By Theorem 1, such a matrix is H7P embeddable if and only if all eigenvalues are positive and satisfy

$$\begin{aligned} &\lambda_{(0,0,0)} = 1, \\ &\lambda_{(0,1,0)}\lambda_{(1,0,0)}\lambda_{(1,1,0)} \geq \lambda_{(0,0,1)}\lambda_{(0,1,1)}\lambda_{(1,0,1)}\lambda_{(1,1,1)}, \\ &\lambda_{(0,0,1)}\lambda_{(1,0,0)}\lambda_{(1,0,1)} \geq \lambda_{(0,1,0)}\lambda_{(0,1,1)}\lambda_{(1,1,0)}\lambda_{(1,1,1)}, \\ &\lambda_{(0,1,1)}\lambda_{(1,0,0)}\lambda_{(1,1,1)} \geq \lambda_{(0,0,1)}\lambda_{(0,1,0)}\lambda_{(1,0,1)}\lambda_{(1,1,0)}, \\ &\lambda_{(0,0,1)}\lambda_{(0,1,0)}\lambda_{(0,1,1)} \geq \lambda_{(1,0,0)}\lambda_{(1,0,1)}\lambda_{(1,1,0)}\lambda_{(1,1,1)}, \\ &\lambda_{(0,1,0)}\lambda_{(1,0,1)}\lambda_{(1,1,1)} \geq \lambda_{(0,0,1)}\lambda_{(0,1,1)}\lambda_{(1,0,0)}\lambda_{(1,1,0)}, \\ &\lambda_{(0,0,1)}\lambda_{(1,1,0)}\lambda_{(1,1,1)} \geq \lambda_{(0,1,0)}\lambda_{(0,1,1)}\lambda_{(1,0,0)}\lambda_{(1,0,1)}, \\ &\lambda_{(0,1,1)}\lambda_{(1,0,1)}\lambda_{(1,1,0)} \geq \lambda_{(0,0,1)}\lambda_{(0,1,0)}\lambda_{(1,0,0)}\lambda_{(1,1,1)}. \end{aligned}$$

5.2 Hachimoji 3-parameter model

The second model we suggest specializes to the Kimura 2-parameter model when restricted to the standard 4-letter DNA. We will call it the *hachimoji 3-parameter (H3P) model*. We recall that in the Kimura 2-parameter model there are three distinct parameters for the rates of mutation: One parameter for a state remaining unchanged, one parameter for transversion from a purine base to a pyrimidine base or vice versa, and one parameter for transition to the other purine or to the other pyrimidine. We say that two bases are of the same type if they are both standard or synthetic bases. In the hachimoji 3-parameter model, there are the following parameters:

- a: the probability of a state remaining unchanged.
- b: the probability of a transversion from a purine base to a pyrimidine base or vice versa.
- c: the probability of a transition to another purine or pyrimidine base of the same type (same type transitions).
- d: the probability of a transition to another purine or pyrimidine base of different type (different type transitions).

The H3P rate and transition matrices have the form

$$P = \begin{pmatrix} a & b & b & c & d & b & b & d \\ b & a & c & b & b & d & d & b \\ b & c & a & b & b & d & d & b \\ c & b & b & a & d & b & b & d \\ d & b & b & d & a & b & b & c \\ b & d & d & b & b & a & c & b \\ b & d & d & b & b & c & a & b \\ d & b & b & d & c & b & b & a \end{pmatrix}. \tag{5.3}$$

The labeling function of this model corresponds to the partition

$$\{\{(0,0,0)\},\{(0,0,1),(0,1,0),(1,0,1),(1,1,0)\},\{(0,1,1)\},\{(1,0,0),(1,1,1)\}\},$$

which is $(\mathcal{G}, L)$-compatible by Table 1.

The eigenvalues of a H3P Markov matrix are

$$\begin{aligned} &w := \lambda_{(0,0,0)} = a + 4b + c + 2d = 1, x := \lambda_{(0,1,1)} = a - 4b + c + 2d, \\ &y := \lambda_{(1,0,0)} = \lambda_{(1,1,1)} = a + c - 2d, z := \lambda_{(0,0,1)} \\ &= \lambda_{(0,1,0)} = \lambda_{(1,0,1)} = \lambda_{(1,1,0)} = a - c. \end{aligned}$$

By Theorem 1, a H3P Markov matrix P is H3P embeddable if and only if the eigenvalues of P satisfy

$$w = 1, 1 \geq x > 0, y > 0, z > 0, x \geq y^2, xy^2 \geq z^4. \tag{5.4}$$

5.3 Hachimoji 1-parameter model

The third model we suggest is the analogue of the Jukes–Cantor model and we will refer to it as *hachimoji 1-parameter (H1P) model*. It is the simplest group-based model associated to the group $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ and it is described by only two distinct parameters for the rates of mutation. The two parameters are for a state remaining the same and a state mutating to any other state. The corresponding labeling function is $(\mathcal{G}, L)$-compatible by Lemma 6. The H1P rate and transition matrices have the form

$$\begin{pmatrix} a & b & b & b & b & b & b & b \\ b & a & b & b & b & b & b & b \\ b & b & a & b & b & b & b & b \\ b & b & b & a & b & b & b & b \\ b & b & b & b & a & b & b & b \\ b & b & b & b & b & a & b & b \\ b & b & b & b & b & b & a & b \\ b & b & b & b & b & b & b & a \end{pmatrix}. \tag{5.5}$$

The eigenvalues of a H1P Markov matrix are $w := \lambda_{(0,0,0)} = 1$ and $x := \lambda_g = a - b$ for $g \neq 0$. By Theorem 1, such a matrix is H1P embeddable if and only if its eigenvalues satisfy

$$w = 1 \quad \text{and} \quad 1 \geq x > 0. \tag{5.6}$$

Remark 4 The same conditions as in (5.6) characterize model embeddability for the general Jukes–Cantor model as defined in Lemma 6. This is also a special instance of a more general result (Baake and Sumner 2020, Corollary 4.7) on equal-input embeddability. If the order of $\mathcal{G}$ is even, then the notion of general embeddability is equivalent to the notion of model embeddability for the general Jukes–Cantor models by Baake and Sumner (2020, Theorem 4.6).

6 Volume

In this section we compute the relative volumes of model embeddable Markov matrices within some meaningful subsets of Markov matrices by taking advantage of the characterisation of embeddability in terms of eigenvalues. The aim of this section is to describe how large the different sets of matrices are compared to each other and provide intuition of how restrictive is the hypothesis of homogeneous continuous-time models.

We will focus on the hachimoji models and the generalization of the Jukes–Cantor model. We will use the following notation:

(i) Δ is the set of all Markov matrices in a model.
(ii) Δ_+ is the subset of matrices in Δ with only positive eigenvalues.
(iii) Δ_{dd} is the subset of diagonally dominant matrices in Δ, i.e. matrices in Δ such that in each row the diagonal entry is greater or equal than the sum of all other entries in the same row.
(iv) Δ_{me} is the subset of model embeddable transition matrices in Δ.

Biologically, the subspace Δ_{dd} of diagonally dominant matrices consists of matrices with probability of not mutating at least as large as the probability of mutating. If a diagonally dominant matrix is embeddable, it has an identifiable rate matrix (Cuthbert 1972, 1973), namely a unique Markov generator, which is crucial for proving the consistency of many phylogenetic reconstruction methods, such as those based on maximum likelihood methods (Casanellas et al. 2020c; Chang 1996). What is more, the set of Markov matrices with positive eigenvalues Δ_+ includes the multiplicative closure of the transition matrices in the continuous-time version of the model (Sumner et al. 2012). We have the inclusions $\Delta_{me} \subseteq \Delta_+ \subseteq \Delta$ and $\Delta_{dd} \subseteq \Delta_+$. The volumes of these spaces are given for the Kimura 3-parameter model in Roca-Lacostena and Fernández-Sánchez (2018, Theorem 4.1), for the Kimura 2-parameter model in Casanellas et al. (2020b, Proposition 5.1] and for the Jukes–Cantor model in Roca-Lacostena and Fernández-Sánchez (2018, Section 4).

The subsets Δ, Δ_+, Δ_{dd}, and Δ_{me} can be described using the parameterization in terms of the entries of the Markov matrix or in terms of their eigenvalues. We parameterize the relevant subsets of Markov matrices in terms of the eigenvalues of the Markov matrices and compute the volumes using these parametrizations. If φ

Table 2 The estimated volume of the set of H7P embeddable matrices using the hit-and-miss Monte Carlo integration with n sample points

n	10^4	10^5	10^6	10^7
$V(\Delta_{me})$	0.0015	0.00197	0.001946	0.0019678
$V(\Delta_{me} \cap \Delta_{dd})$	0.0008	0.00084	0.00085	0.0008271

denotes the bijection from the set of entries of a Markov matrix in a particular model to the set of its eigenvalues and the matrix $J(\varphi)$ denotes the Jacobian matrix of the map φ, then the volume of any subset in the parametrization using entries of a Markov matrix will be $|det(J(\varphi))|$ times the volume in the parameterization using eigenvalues. Since the determinant of this Jacobian is constant for each of the three models we consider, the relative volumes of the set of model embeddable Markov matrices will not depend on the parameterization chosen.

Proposition 1 *For the hachimoji 7-parameter model, consider* Δ, Δ_+, *and* Δ_{dd} *as subsets of* $\mathbb{R}^7$ *parameterized by* $\lambda_{(0,0,1)}, \ldots, \lambda_{(1,1,1)}$, *the eigenvalues of a H7P Markov matrix. Then: (i)* $V(\Delta) = \frac{256}{315}$; *(ii)* $V(\Delta_+) = \frac{5}{144}$; *(iii)*$V(\Delta_{dd}) = \frac{2}{315}$.

Proof The entries of a H7P Markov matrix (5.2) are determined by a vector (a, b, c, d, e, f, g, h). The entries of this vector can be expressed in terms of the eigenvalues as

$$\begin{aligned}&\big(a, b, c, d, e, f, g, h\big)^T \\ &\quad = K^{-1}\left(1, \lambda_{(0,0,1)}, \lambda_{(0,1,0)}, \lambda_{(0,1,1)}, \lambda_{(1,0,0)}, \lambda_{(1,0,1)}, \lambda_{(1,1,0)}, \lambda_{(1,1,1)}\right)^T,\end{aligned}$$

where K is the discrete Fourier transform matrix (5.1). In terms of the entries or the eigenvalues of a H7P Markov matrix, the relevant subsets in this model are given by:

$$\begin{aligned}&\Delta = \{(a, b, c, d, e, f, g, h) \in \mathbb{R}^8 : a + b + c + d + e + f + g + h = 1,\\ &a, b, c, d, e, f, g, h \geq 0\},\\ &\Delta_+ = \{P \in \Delta : \lambda_{(0,0,1)}, \lambda_{(0,1,0)}, \lambda_{(0,1,1)}, \lambda_{(1,0,0)}, \lambda_{(1,0,1)}, \lambda_{(1,1,0)}, \lambda_{(1,1,1)} > 0\},\\ &\Delta_{dd} = \{P \in \Delta : a \geq b + c + d + e + f + g + h\}, \text{ and}\end{aligned}$$

Δ_{me} is given by one equation and seven inequalities presented in Sect. 5.1. Expressing all conditions defining Δ, Δ_+, and Δ_{dd} in terms of the eigenvalues $\lambda_{(0,0,1)}, \ldots, \lambda_{(1,1,1)}$ allows us to compute volumes of these sets using `Polymake` (Gawrilow and Joswig 2000).

We are not able to compute the volume of the subspace of the H7P embeddable Markov matrices exactly. Instead we estimate the volume using the hit-and-miss Monte Carlo integration method (Hammersley 2013) implemented in `Mathematica`. Table 2 summarizes the volume for various number of sample points. Table 3 gives relative volumes for the relevant sets.

Table 3 The relative volumes for the hachimoji 7-parameter model

	Δ	Δ_+	Δ_{dd}
$\frac{V(\cdot)}{V(\Delta)}$	1	$\frac{175}{4096} = 0.042724609375$	$\frac{1}{128} = 0.0078125$
$\frac{V(\Delta_{me} \cap \cdot)}{V(\cdot)}$	≈ 0.00239	≈ 0.056045	≈ 0.13388

The volumes of Δ_{me} and $\Delta_{me} \cap \Delta_+$ are estimated using Monte Carlo integration with 10^6 sample points

Table 4 The relative volumes for the hachimoji 3-parameter model

	Δ	Δ_+	Δ_{dd}
$\frac{V(\cdot)}{V(\Delta)}$	1	$\frac{21}{64} = 0.328125$	$\frac{1}{8} = 0.125$
$\frac{V(\Delta_{me} \cap \cdot)}{V(\cdot)}$	$\frac{1}{4} = 0.25$	$\frac{16}{21} \approx 0.76190$	≈ 0.82040

Proposition 2 *For the hachimoji 3-parameter model, consider Δ, Δ_+, Δ_{dd}, and Δ_{me} as subsets of $\mathbb{R}^3$ parameterized by x, y, z, the eigenvalues of a H3P Markov matrix. Then: (i) $V(\Delta) = \frac{4}{3}$; (ii) $V(\Delta_+) = \frac{7}{16}$; (iii) $V(\Delta_{dd}) = \frac{1}{6}$; (iv) $V(\Delta_{me}) = \frac{1}{3}$; (v) $V(\Delta_{me} \cap \Delta_{dd}) \approx 0.136733$.*

Proof The entries of a H3P Markov matrix as in (5.3) can be expressed in terms of the eigenvalues as

$$a = \frac{1 + x + 2y + 4z}{8}, \quad b = \frac{1 - x}{8},$$
$$c = \frac{1 + x + 2y - 4z}{8}, \quad d = \frac{1 + x - 2y}{8}.$$

Expressing all conditions defining Δ, Δ_+, Δ_{dd}, and Δ_{me} in terms of x, y, z allows us to use the `Integrate` command in `Mathematica` to compute the desired volumes. For $V(\Delta_{me} \cap \Delta_{dd})$ we used the numerical integration command `NIntegrate`.

The sets Δ_{me}, Δ_+ and Δ for the hachimoji 3-parameter model are depicted in Fig. 1. The relative volumes of relevant sets are given in Table 4.

Finally, we discuss the generalization of the Jukes–Cantor model which includes the hachimoji 1-parameter model. Let $\mathcal{G}$ be a finite abelian group of order n and $L : \mathcal{G} \to \{0, 1\}$ a labeling function such that and $L(0) = 0$ and $L(g) = 1$ for $g \neq 0$. In Lemma 6 we proved that L is a $\mathcal{G}$-compatible labeling. In the general Jukes–Cantor model, the transition matrix P corresponding to this labeling is has the form

$$P_{ij} = \begin{cases} a, & i = j \\ b, & i \neq j \end{cases}.$$

Since P is a Markov matrix, then $a = 1 - (n-1)b$, and thus P is parameterized by b.

Proposition 3 *For the general Jukes–Cantor model, consider Δ, Δ_+, Δ_{dd}, Δ_{me} as subsets of $\mathbb{R}$ parameterized by b, the off-diagonal element of the Markov matrix. Then:*

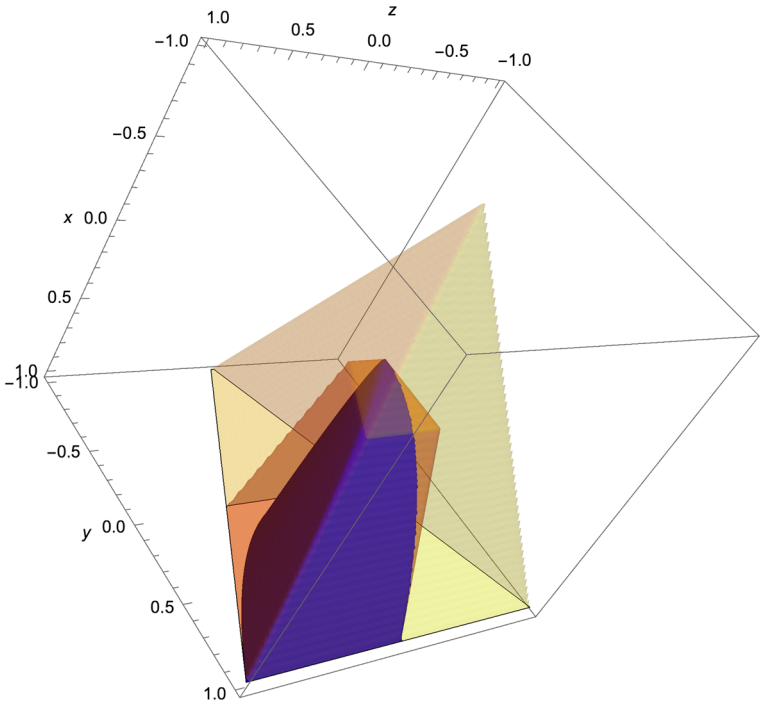

Fig. 1 The sets Δ_{me}, Δ_+ and Δ for the hachimoji 3-parameter model. The sets Δ_+ and Δ are polytopes; the set Δ_{me} is a semialgebraic set

(i) $\Delta = [0, \frac{1}{n-1}]$*; (ii)* $\Delta_+ = [0, \frac{1}{n})$*; (iii)* $\Delta_{dd} = [0, \frac{1}{2(n-1)}]$*; (iv)* $\Delta_{me} = [0, \frac{1}{n})$*; (v)* $\Delta_{me} \cap \Delta_{dd} = [0, \frac{1}{2(n-1)}]$.

Proof The Markov matrix P has eigenvalues 1 with multiplicity 1 and $a - b = 1 - nb$ with multiplicity $n - 1$. Hence

(i) $\Delta = \{b \in \mathbb{R} : a = 1 - (n-1)b \geq 0, b \geq 0\} = [0, \frac{1}{n-1}]$.

(ii) $\Delta_+ = \{b \in \mathbb{R} : a = 1 - (n-1)b \geq 0, b \geq 0, 1 - nb > 0\} = [0, \frac{1}{n})$.

(iii) $\Delta_{dd} = \{b \in \mathbb{R} : a = 1 - (n-1)b \geq 0, b \geq 0, 1 - (n-1)b \geq (n-1)b\} = [0, \frac{1}{2(n-1)}]$.

(iv) By Remark 4, a Markov matrix is general Jukes–Cantor embeddable if and only if the eigenvalue $1 - nb$ satisfies $1 \geq 1 - nb > 0$. Since $1 \geq 1 - nb$ necessarily holds for any Markov matrix, we have $\Delta_{me} = \Delta_+$.

(v) Since $\Delta_{dd} \subseteq \Delta_+ = \Delta_{me}$, then $\Delta_{me} \cap \Delta_{dd} = \Delta_{dd}$.

The relative volumes of relevant sets for the general Jukes–Cantor model are presented in Table 5. Proposition 3 gives for the hachimoji 1-parameter model (i) $\Delta = [0, \frac{1}{7}]$; (ii) $\Delta_+ = [0, \frac{1}{8})$; (iii)$\Delta_{dd} = [0, \frac{1}{14}]$; (iv) $\Delta_{me} = [0, \frac{1}{8})$; (v)$\Delta_{me} \cap \Delta_{dd} = [0, \frac{1}{14}]$.

Table 5 The relative volumes for the general Jukes–Cantor model

	Δ	Δ_+	Δ_{dd}
$\frac{V(\cdot)}{V(\Delta)}$	1	$\frac{n-1}{n}$	$\frac{1}{2}$
$\frac{V(\Delta_{me} \cap \cdot)}{V(\cdot)}$	$\frac{n-1}{n}$	1	1

7 Conclusion

When modelling sequence evolution we often adopt several simplifying assumptions, which make the statistical problems tractable. The commonly used Markov models depend on the assumption that sites evolve independently following a Markov process. The Markov chain is often assumed to be homogeneous continuous-time, that is the transition probabilities are independent of the time. This means that the instantaneous rates of substitution at any time are fixed and usually displayed as the entries of rate matrices. If an evolutionary process is not homogeneous, then one can multiply transition matrices of short homogeneous processes. The resulting matrix is not necessarily embeddable, but if it is, then the inhomogeneous process can be approximated by a homogeneous one.

In this paper we provide necessary and sufficient conditions for model-embeddability of $n \times n$ symmetric group-based substitution models, which include the well known Cavender–Farris–Neyman, Jukes–Cantor, Kimura-2 and Kimura-3 parameter models for DNA. We fully characterize those embeddable $n \times n$ stochastic matrices following a symmetric group-based model structure whose Markov generators also satisfy the constraints of the model, which we refer to as model embeddability.

A novel application of our main result is the characterization of model embeddability for three group-based models for the hachimoji DNA, a synthetic genetic system with eight building blocks. For these models we also compute the relevant volumes of model embeddable matrices within other relevant sets of Markov matrices. These computations show how restrictive is the hypothesis of a particular hachimoji time-continuous group-based model.

In this article we have considered symmetric group-based models. The importance of the symmetricity assumption is that it guarantees that the eigenvalues of rate and transition matrices of a group-based model are real. We use this property in the proof of Theorem 1. A future research question is to explore whether this approach can be extended to group-based models that are not symmetric.

Acknowledgements An initial version of the main result of the present manuscript appeared in the preprint arXiv:1705.09228. Following the advice of referees, we divided the preprint into two manuscripts. The present manuscript builds on the section on the embedding problem in the earlier preprint. Most of the preprint arXiv:1705.09228 appeared in Bulletin of Mathematical Biology under the title "Maximum Likelihood Estimation of Symmetric Group-Based Models via Numerical Algebraic Geometry".

Funding Open access funding provided by Aalto University. Dimitra Kosta was partially supported by a Royal Society Dorothy Hodgkin Research Fellowship. Kaie Kubjas was partially supported by the Academy of Finland Grant 323416.

A Friendly labeling functions

Besides $\mathcal{G}$-compatible labeling functions, there is another class of labeling functions which has been studied in the literature. They are called *friendly labeling functions* and were introduced by Sturmfels and Sullivant (2005). Friendly labelings are useful in determining phylogenetic invariants for group-based models on evolutionary trees. In particular, a friendly labeling guarantees that if a particular labeling comes from an assignment of group elements, then any choice of a group element to one particular edge which is consistent with the labeling can be extended to an assignment that is consistent with labeling on all edges of the claw tree.

Definition 3 Let $\mathcal{G}$ be a finite abelian group and $L : \mathcal{G} \to \mathcal{L}$ a labeling function. Let $n \in \mathbb{N}$ and $Z := \{g \in \mathcal{G}^n : g_n = \sum_{i=1}^{n-1} g_i\}$. Define the map $\tilde{L} : Z \subseteq \mathcal{G}^n \to \mathcal{L}^n$ to be the induced labeling function on $Z \subseteq \mathcal{G}^n$. The labeling function L is said to be *n-friendly* if for every $l \in \tilde{L}(Z)$ and $i = 1, 2, \cdots, n$, we have $\pi_i(\tilde{L}^{-1}(l)) = L^{-1}(\pi_i(l))$. Here, π_i denotes the projection to the i-th component. Furthermore, the labeling function L is said to be *friendly* if it is n-friendly for all $n \geq 3$.

By Sturmfels and Sullivant (2005, Lemma 11), to check whether a labeling function is friendly, it is enough to check that the labeling is 3-friendly.

Example 6 (Sturmfels and Sullivant 2005, Example 9) Let $\mathcal{G} = \mathbb{Z}_4$ and $L : \mathcal{G} \to \{0, 1, 2\}$ such that

$$L(0) = 0, L(1) = 1, L(2) = L(3) = 2.$$

Then L is not friendly labeling because $L^{-1}(\pi_3((1, 1, 2))) = \{2, 3\}$ while $\pi_3(\tilde{L}^{-1}(1, 1, 2)) = \pi_3((1, 1, 2)) = \{2\}$.

Table 6 summarizes all friendly labelings for abelian groups of order n, where $2 \leq n \leq 8$. In the table, two group elements receive the same label if they belong to the same subset in a partition of $\mathcal{G}$. In the table we do not include the friendly labelings for $\mathbb{Z}_2 \times \mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, since there are too many of them.

The next example shows that there are friendly labelings that are not $\mathcal{G}$-compatible.

Example 7 Let $\mathcal{G}$ be a finite abelian group and $L : \mathcal{G} \to \{0\}$ the labeling function defined by $L(g) = 0$ for all $g \in \mathcal{G}$. By Lemma 5, the labeling function L is not $\mathcal{G}$-compatible. However, it is a friendly labeling, since $\pi_i(\tilde{L}^{-1}((0, 0, 0))) = \pi_i(Z) = \mathcal{G}$ and $L^{-1}(\pi_i((0, 0, 0))) = L^{-1}(0) = \mathcal{G}$.

Table 6 Friendly labelings for abelian group of order $n \leq 8$

n	Group	Friendly labelings
2	$\mathbb{Z}_2$	{{0,1}},{{0},{1}}
3	$\mathbb{Z}_3$	{{0,1,2}},{{0},{1,2}},{{0},{1},{2}}
4	$\mathbb{Z}_4$	{{0,1,2,3}},{{0},{1,2,3}},{{0,2},{1,3}},{{0},{1,3},{2}},{{0},{1},{2},{3}}
4	$\mathbb{Z}_2 \times \mathbb{Z}_2$	{{(0,0),(0,1),(1,0),(1,1)}},{{(0,0)},{(0,1),(1,0),(1,1)}}, {{(0,0),(0,1)},{(1,0),(1,1)}},{{(0,0),(1,1)},{(0,1),(1,0)}}, {{(0,0),(1,0)},{(0,1),(1,1)}},{{(0,0)},{(0,1)},{(1,0),(1,1)}}, {{(0,0)},{(0,1),(1,0)},{(1,1)}},{{(0,0)},{(0,1),(1,1)},{(1,0)}}, {{(0,0)},{(0,1)},{(1,0)},{(1,1)}}
5	$\mathbb{Z}_5$	{{0,1,2,3,4}},{{0},{1,2,3,4}},{{0},{1,4},{2,3}},{{0},{1},{2},{3},{4}}
6	$\mathbb{Z}_2 \times \mathbb{Z}_3$	{{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2)}},{{(0,0)},{(0,1),(0,2),(1,0),(1,1),(1,2)}}, {{(0,0),(0,1),(0,2)},{(1,0),(1,1),(1,2)}},{{(0,0),(1,0)},{(0,1),(0,2),(1,1),(1,2)}}, {{(0,0)},{(0,1),(0,2)},{(1,0),(1,1),(1,2)}},{{(0,0)},{(0,1),(0,2),(1,1),(1,2)}, {(1,0)}},{{(0,0),(1,0)},{(0,1),(1,1)},{(0,2),(1,2)}}, {{(0,0)},{(0,1)},{(0,2)},{(1,0),(1,1),(1,2)}}, {{(0,0)},{(0,1),(0,2)},{(1,0)},{(1,1),(1,2)}},{{(0,0)},{(0,1),(1,1)},{(0,2),(1,2)}, {(1,0)}}, {{(0,0)},{(0,1)},{(0,2)},{(1,0)},{(1,1)},{(1,2)}}
7	$\mathbb{Z}_7$	{{0,1,2,3,4,5,6}},{{0},{1,2,3,4,5,6}},{{0},{1,2,4},{3,5,6}}, {{0},{1,6},{2,5},{3,4}},{{0},{1},{2},{3},{4},{5},{6}}
8	$\mathbb{Z}_8$	{{0,1,2,3,4,5,6,7}},{{0},{1,2,3,4,5,6,7},{{0,4},{1,2,3,5,6,7}}, {{0,2,4,6},{1,3,5,7}},{{0},{1,2,3,5,6,7},{4}},{{0},{1,2,6,7},{3,4,5}}, {{0},{1,4,7},{2,3,5,6}},{{0},{1,3,5,7},{2,4,6}},{{0,4},{1,3,5,7},{2,6}}, {{0},{1,3,5,7},{2,6},{4}},{{0,4},{1,5},{2,6},{3,7}},{{0},{1,3,5,7},{2},{4}, {6}},{{0},{1,3}{2,6},{4},{5,7}},{{0},{1,7},{2,6},{3,5},{4}},{{0},{1,5},{2,6}, {3,7},{4}},{{0},{1,5},{2},{3,7},{4},{6}},{{0},{1},{2},{3},{4},{5},{6},{7}}

Computations for abelian groups of order at most eight demonstrate that every symmetric $\mathcal{G}$-compatible labeling is a friendly labeling. It is left open, if the same is true for any finite abelian group.

Question 1 *Given a finite abelian group $\mathcal{G}$, is the set of all symmetric $\mathcal{G}$-compatible labelings strictly contained in the set of all friendly labelings?*

References

Allman E, Rhodes J (2003) Phylogenetic invariants for the general Markov model of sequence mutation. Math Biosci 186:133–144

Baake M, Sumner J (2020) Notes on Markov embedding. Linear Algebra Appl 594:262–299

Baños H, Bushek N, Davidson R, Gross E, Harris PE, Krone R, Long C, Stewart A, Walker R (2016) Phylogenetic trees. arXiv:1611.05805

Carette P (1995) Characterizations of embeddable 3×3 stochastic matrices with a negative eigenvalue. N Y J Math 1:120–129

Casanellas M, Fernández-Sánchez J (2007) Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees. Mol Biol Evol 24(1):288–293

Casanellas M, Fernández-Sánchez J, Roca-Lacostena J (2020a) Embeddability and rate identifiability of Kimura 2-parameter matrices. J Math Biol 80(4):995–1019

Casanellas M, Fernàndez-Srnànchez J, Roca-Lacostena J (2020b) The embedding problem for Markov matrices. 2005.00818

Casanellas M, Petrović S, Uhler C (2020c) Algebraic statistics in practice: applications to networks. Annu Rev Stat Appl 7:227–250
Cavender J (1978) Taxonomy with confidence. Math Biosci 40:271–280
Cavender J, Felsenstein J (1987) Invariants of phylogenies in a simple case with discrete states. Classification 4:57–71
Chang JT (1996) Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci 137(1):51–73
Chen Y, Chen J (2011) On the imbedding problem for three-state time homogeneous Markov chains with coinciding negative eigenvalues. J Theor Probab 24:928–938
Cuthbert RJ (1972) On uniqueness of the logarithm for Markov semi-groups. J Lond Math Soc 2(4):623–630
Cuthbert RJ (1973) The logarithm function for finite-state Markov semi-groups. J Lond Math Soc 2(3):524–532
Davies EB (2010) Embeddable Markov matrices. Electron J Probab 15:1474–1486
Elfving G (1937) Zur theorie der Markoffschen ketten. Acta Societatis Scientiarum FennicæNova Series A 2(8):17
Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. Ann Stat 21:355–377
Farris JS (1973) A probability model for inferring evolutionary trees. Syst Zool 22(3):250–256
Felsenstein J (2003) Inferring phylogenies. Sinauer Associates Inc, Publishers, Sunderland
Gawrilow E, Joswig M (2000) `polymake`: a framework for analyzing convex polytopes. In: Polytopes—combinatorics and computation (Oberwolfach, 1997), DMV Sem., vol 29, Birkhäuser, Basel, pp 43–73
Gross E, Long C (2018) Distinguishing phylogenetic networks. SIAM J Appl Algebra Geom 2(1):72–93
Hammersley J (2013) Monte Carlo methods. Springer, Berlin
Hoshika S, Leal NA, Kim MJ, Kim MS, Karalkar NB, Kim HJ, Bates AM, Watkins NE, SantaLucia HA, Meyer AJ et al (2019) Hachimoji DNA and RNA: a genetic system with eight building blocks. Science 363(6429):884–887
Israel RB, Rosenthal JS, Wei JZ (2001) Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. Math Finance 11(2):245–265
Jia C (2016) A solution to the reversible embedding problem for finite Markov chains. Stat Probab Lett 116:122–130
Johansen S (1974) Some results on the imbedding problem for finite Markov chains. J Lond Math Soc 2:345–351
Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro H (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132
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 USA 78(1):454–458
Kingman JFC (1962) The imbedding problem for finite Markov chains. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 1(1):14–24
Lake JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191
Luong B (2009) Fourier analysis on finite Abelian groups. Springer, Berlin
Matsen FA (2008) Fourier transform inequalities for phylogenetic trees. IEEE/ACM Trans Comput Biol Bioinform 6(1):89–95
Michałek M (2011) Geometry of phylogenetic group-based models. J Algebra 339(1):339–356
Neyman J (1971) Molecular studies of evolution: a source of novel statistical problems. In: Gupta SS, Yackel J (eds) Stat Decis Theory Relat Top. Academic Press, New York, pp 1–27
Pachter L, Sturmfels B (2004) Tropical geometry of statistical models. Proc Natl Acad Sci USA 101(46):16132–16137
Pachter L, Sturmfels B (2005) Algebraic statistics for computational biology, vol 13. Cambridge University Press, Cambridge
Roca-Lacostena J, Fernández-Sánchez J (2018) Embeddability of Kimura 3ST Markov matrices. J Theor Biol 445:128–135
Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12(2):204–228
Sumner JG, Fernández-Sánchez J, Jarvis PD (2012) Lie Markov models. J Theor Biol 298:16–31

Verbyla KL, Von Bing Yap AP, Shao Y, Huttley GA (2013) The embedding problem for Markov models of nucleotide substitution. PLoS ONE 8(7):e69187