Journal of Theoretical Biology 445 (2018) 128–135

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Embeddability of Kimura 3ST Markov matrices

Jordi Roca-Lacostena, Jesús Fernández-Sánchez*

Departament de Matemátiques, Universitat Politécnica de Catalunya, Spain

ARTICLE INFO

Article history:
Received 5 May 2017
Revised 24 January 2018
Accepted 5 February 2018
Available online 17 February 2018

Keywords:
Markov matrix
Markov generator
Eigenvalues
Evolutionary model
Embeddability

ABSTRACT

In this note, we characterize the embeddability of generic Kimura 3ST Markov matrices in terms of their eigenvalues. As a consequence, we are able to compute the volume of such matrices relative to the volume of all Markov matrices within the model. We also provide examples showing that, in general, mutation rates are not identifiable from substitution probabilities. These examples also illustrate that symmetries between mutation probabilities do not necessarily arise from symmetries between the corresponding mutation rates.

1. Introduction

Genomic data expressed by means of sequence alignments is widely used to infer phylogenetic relationships between species. Substitution models are used to describe the evolutionary process that leads from one DNA sequence to another. These models are usually given in terms of a family of Markov matrices with a prescribed structure. The entries of these matrices represent the conditional probabilities of nucleotide substitution between one sequence and the other, and can be obtained either by counting the relative frequencies of these substitutions or fitting the parameters of the model using maximum likelihood. Usually, the structure imposed by the model is motivated by some biological / biochemical properties observed (e.g. the Kimura 3ST model Kimura, 1981) or some computational / mathematical convenient assumptions to deal with the model (e.g. the GTR model Tavaré, 1986 or Lie Markov models Sumner et al., 2012). Moreover, evolution is usually modelled by means of Markov chains, together with the additional assumption that all sites in the sequences evolve independently and according to the same probabilities.

A general approach in modelling evolution corresponds to regarding time as a continuous variable where substitution events always happen at the same rate, which remains constant throughout the whole evolutionary process. This leads to the homogeneous continuous-time substitution models, where only Markov matrices that are the exponential of a rate matrix are considered. Clearly this is used as an approximation to biological reality where it is well known that transition rates vary over time (Ho et al., 2005; 2007) and also among the different branches of the phylogenetic tree (Lockhart et al., 1998). However, given the bias / variance compensation of the statistical analysis (Burnham and Anderson, 2002), modelling phylogenetic evolution as a non-homogeneous process is not statistically feasible in practice (cf. Sumner et al., 2012).

A different approach appears when one regards the evolutionary process as a whole and only takes into account the conditional probabilities between the original and the final sequences, without caring about rates of mutation. When these probabilities are taken as the parameters of the model, we deal with the so-called *algebraic* models.[1] Algebraic models have been used in a number of theoretical papers, including Allman and Rhodes (2008); Casanellas and Fernández-Sánchez (2010); Draisma and Kuttler (2008); Sturmfels and Sullivant (2005).

If one attempts to connect both approaches, a natural question is to decide whether a given Markov matrix is the exponential of some rate matrix, whose entries would be some kind of average of the rates involved throughout the evolutionary process. In this case, we say the matrix is *embeddable* and this question is known in the literature as the *embedding problem* for Markov matrices. An easier version of this problem is to decide whether the rate matrices associated to the embeddable matrices of a particular (algebraic) model $\mathcal{M}$ should keep the same symmetries as the model ($\mathcal{M}$*-embeddability*, see definition in Section 2.2). The embedding

* Corresponding author.
E-mail addresses: jordi.roca.lacostena@upc.edu (J. Roca-Lacostena), jesus.fernandez.sanchez@upc.edu (J. Fernández-Sánchez).

[1] Here, “algebraic” refers to the fact that the probabilities of pattern observation at the leaves of a phylogenetic tree evolving under these models are given by algebraic expressions (only sums and products) in terms of the parameters of the model.

https://doi.org/10.1016/j.jtbi.2018.02.005
0022-5193/© 2018 Elsevier Ltd. All rights reserved.

problem is relevant even if restricted to continuous-time models since it is not true in general that the product of embeddable matrices is necessarily embeddable (indeed, the Baker–Campbell–Hausdorff formula Campbell, 1897 leads to ask whether some series of matrices is convergent or not, which is not always true Blanes and Casas, 2004). These questions are closely related to the problem of the multiplicative closure of continuous-time models, namely whether the product of matrices $e^{Q_1}e^{Q_2}$ where Q_1 and Q_2 are rate matrices in one particular (continuous-time) model can be obtained as some e^{Q} for some rate matrix Q in *the same model*. After Sumner et al. (2012) and Sumner (2017), it is known that there are popular models which are not multiplicatively closed, notably including the GTR model and the HKY model.

The reader is referred to Davies (2010) for a nice overview of the embedding problem from a mathematical point of view. In a more biological and applied setting, the paper by Verbyla et al. (2013) deals with the possible consequences for phylogenetic inference. Also, the paper Sumner et al. (2012) and the more recent paper Woodhams et al. (2017) deal with the incidental question of how the lack of (multiplicative) closure in substitution models have consequences for the phylogenetic analysis of data.

In this paper, we deal with the embedding problem from a theoretical perspective. The main goal is to obtain a characterization for the embeddability of generic matrices of the Kimura 3ST model (Kimura, 1981). From our results, we will be able to compute the whole volume of embeddable Kimura 3ST matrices and compare it with the volume of the whole space of Kimura 3ST Markov matrices. At the same time, we provide a number of examples showing matrices that are embeddable but for which the mutation rates are not identifiable or do not keep the same structure of the model. The recent paper Kosta and Kubjas (2017) deals with the similar question of characterizing embeddable matrices of symmetric group-based phylogenetic models, but focusing on the existence of rate matrices strictly in the model.

The organization of the paper is as follows. In Section 2, we recall some definitions and basic facts concerning the embedding problem and the Kimura 3ST model. Here, we also show that any embeddable matrix is biologically relevant since it can be seen as the transition matrix of a concatenation of *realistic* evolutionary processes (“realistic” here means a process whose transition matrix is close to the identity matrix, see Theorem 2.2). In Section 3, we prove the main theorem which characterizes under the (generic) assumption of having different eigenvalues the Kimura 3ST embeddable matrices in terms of inequalities to be satisfied by the eigenvalues. We devote as well some attention to the case of matrices with repeated eigenvalues as they present certain situations that may be interesting from a theoretical and applied point of view. Namely, these matrices show that the identifiability of the mutation rates is not a generic property for the Kimura 2ST model or the Jukes–Cantor model, as well as that there are embeddable matrices with rate matrices that do not keep the same symmetries of the model (see Theorem 3.2). As a consequence of the characterization mentioned above, in Section 4 we are able to compute the volume of embeddable matrices and compare it to the volume of all Kimura 3ST Markov matrices. Finally, Section 5 discusses implications and possibilities for future work.

2. Preliminaries

2.1. Embedding problem of Markov matrices

We denote by $M_k(\mathbb{K})$ the space of all square k-matrices with entries in a field $\mathbb{K}$, where $\mathbb{K}$ is $\mathbb{R}$ or $\mathbb{C}$. Given a matrix $A \in M_k(\mathbb{K})$, we say that $B \in M_k(\mathbb{K})$ is a *logarithm* of A if $e^B = A$, where the exponential of a matrix is defined as

$$e^X = \sum_{n\geq 0} \frac{X^n}{n!}.$$

A classical result states that $det(e^X) = e^{tr(X)}$, so the determinant of any matrix of the form e^X is never 0. Given a non-negative complex number $x \in \mathbb{C}\setminus\mathbb{R}^-$, we will denote by $log(x)$ its *principal logarithm*, that is, the only logarithm of x that lies in the strip $\{z \mid -\pi < Im(z) < \pi\}$. Although the exponential map of matrices is not injective, it is known that if A is a matrix with no negative eigenvalues, there is a unique logarithm X of A all of whose eigenvalues are given by the principal logarithm of the eigenvalues of A (Theorem 1.31 of Higham, 2008). We will refer to this as the *principal logarithm of A* and we will denote it by $Log(A)$. In the particular case where the matrix A is diagonalizable, $A = SDS^{-1}$ then $Log(A) = SLog(D)S^{-1}$, where $Log(D)$ is the diagonal matrix with diagonal entries equal to the principal logarithm of the eigenvalues of A.

Definition 2.1. A matrix $M \in M_k(\mathbb{R})$ is said to be a *Markov matrix* if all the entries are non-negative and the rows sum to one. A matrix $Q \in M_k(\mathbb{R})$ is said to be a *rate matrix* if all the non-diagonal entries are non-negative and the rows sum to zero.

If Q is a rate matrix, it is well-known that $e^{tQ} = \sum_{n\geq 0}\frac{t^nQ^n}{n!}$ is a Markov matrix for all $t \geq 0$. That is why rate matrices are also referred as *Markov generators* (Davies, 2010). However, not every Markov matrix can be obtained in this way. A Markov matrix M is said to be *embeddable* if $M = e^Q$ for some rate matrix Q. The *embedding problem* attempts to decide which (Markov) matrices are embeddable, that is, which matrices can be written as $M = e^Q$, where Q is a rate matrix. We would like to point out that every embeddable matrix can be obtained as the substitution matrix of a long-running biologically realistic Markov process. Namely,

Theorem 2.2. *Every embeddable matrix is the product of embeddable matrices close to the identity matrix.*

Proof. Assume that M is an embeddable Markov matrix: $M = e^Q$. Clearly, $Q_n := \frac{1}{n}Q$ is still a rate matrix for any $n \geq 1$, so $M = (e^{Q_n})^n$ appears as the n-th power of a Markov matrix. Moreover, since

$$\lim_{n\to\infty} e^{Q_n} = e^{\lim_{n\to\infty} Q_n} = e^{(0)} = Id,$$

we can take n big enough so that e^{Q_n} is as close to Id as wanted. □

2.2. Kimura models

In this work we deal with the substitution model introduced by Kimura (1981). The Kimura 3ST model assigns three parameters to different type of substitutions: one parameter for transitions, i.e. substitutions between purines (A ↔ G) or pyrimidines (C ↔ T), and two parameters for transversions, i.e. substitutions that change the type of nucleotide: from purine to pyrimidine or vice versa. Ordering the set of nucleotides as A, G, C, T, the Markov matrices within the model are described by the following structure:

Definition 2.3. A matrix $M \in M_4(\mathbb{C})$ is *Kimura 3ST* (is K3 or has K3 form, for short) if it has the following structure:

$$M = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}. \tag{1}$$

For ease of reading we will use the notation $M = K(a, b, c, d)$ to denote a matrix with the structure in (1).