Journal of Theoretical Biology 298 (2012) 16–31

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Lie Markov models

J.G. Sumner [a,*,1], J. Fernández-Sánchez [b], P.D. Jarvis [a,2]

[a] School of Mathematics and Physics, University of Tasmania, Australia
[b] Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Spain

ARTICLE INFO

Article history:
Received 29 July 2011
Received in revised form
15 December 2011
Accepted 16 December 2011
Available online 29 December 2011

Keywords:
Phylogenetics
Lie algebras
Representation theory
Symmetry
Markov chains

ABSTRACT

Recent work has discussed the importance of multiplicative closure for the Markov models used in phylogenetics. For continuous-time Markov chains, a sufficient condition for multiplicative closure of a model class is ensured by demanding that the set of rate-matrices belonging to the model class form a Lie algebra. It is the case that some well-known Markov models *do* form Lie algebras and we refer to such models as "Lie Markov models". However it is also the case that some other well-known Markov models unequivocally *do not* form Lie algebras (GTR being the most conspicuous example).

In this paper, we will discuss how to generate Lie Markov models by demanding that the models have certain symmetries under nucleotide permutations. We show that the Lie Markov models include, and hence provide a unifying concept for, "group-based" and "equivariant" models. For each of two and four character states, the full list of Lie Markov models with maximal symmetry is presented and shown to include interesting examples that are neither group-based nor equivariant. We also argue that our scheme is pleasing in the context of applied phylogenetics, as, for a given symmetry of nucleotide substitution, it provides a natural hierarchy of models with increasing number of parameters. We also note that our methods are applicable to any application of continuous-time Markov chains beyond the initial motivations we take from phylogenetics.

1. Introduction

Continuous-time Markov chains are fundamental to the implementation of, and philosophy behind, many phylogenetic methods. Likelihood and Bayesian phylogenetic methods usually proceed by attempting to fit a single "rate-matrix" globally across a proposed evolutionary tree history (see, for example, Chapters 2 and 3 of Gascuel, 2005). These rate-matrices are chosen from some restricted class or "model" that is defined by a certain set of constraints on the elements of a generic rate-matrix. These constraints define a set of free parameters that usually correspond to unknown evolutionary quantities such as base composition, mutation rates and the timing of speciation events (these last two are often by necessity confounded together simply as "edge lengths"). Even in phylogenetic distance methods, it is usually the case that the theoretical justification of a given distance estimator is taken from a continuous-time Markov model (for example the general Markov model for the "log-det" (Steel, 1994) distance or the HKY distance taken from its corresponding model, Felsenstein, 2004).

A *homogeneous* Markov chain satisfies the condition that the probability transition rates are constant in time. In the phylogenetic context this means that the rates are unchanged throughout evolutionary history. Of course, this is used as an approximation to biological reality where it is well documented that transition rates are not only time-dependent (Ho et al., 2005, 2007), but also vary across the different lineages of the evolutionary tree (Lockhart et al., 1998). Methods to cope with these issues have been explored by various authors: Tuffley and Steel (1997) proposed the "covarion" model where a switching process allows sites to alternate between "on" and "off" states. Drummond et al. (2006) proposed a method that introduces an overall scaling factor for the transition rates that is sampled randomly (at branching events, for example), and the methods presented in Whelan (2008) are more general still with a switching process that allows for alteration of individual rates. The simulation package discussed in Fletcher and Yang (2009) provides further evidence that these issues are of ongoing importance to phylogenetic analysis.

Our philosophy is to remain agnostic as to whether evolutionary rates have changed in the past or, indeed, whether it is possible to statistically detect this change via analysis of present day molecular data. We follow an approach that allows for the biological possibility that there is likely to have been a smooth (or even abrupt) change of each *individual* transition rate independently occurring across the evolutionary tree (and not necessarily restricted to branching events).

* Corresponding author.
E-mail address: jsumner@utas.edu.au (J.G. Sumner).
[1] ARC Research Fellow.
[2] Alexander von Humboldt Fellow.

0022-5193/$ - see front matter Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.12.017

This discussion leads naturally to confronting the possibility (at least theoretically) that the phylogenetic process is not homogeneous and is more accurately modeled as an *inhomogeneous* continuous-time Markov chain; where the rate-matrix is far from constant and ultimately is allowed to vary, smoothly or otherwise, as a function of edge length parameters of the evolutionary tree.

Of course, given the bias/variance tradeoff of statistical analysis (Burnham and Anderson, 2002), modeling phylogenetic evolution as a inhomogeneous process is statistically implausible in practice (we would effectively be replacing a small number of parameters by an infinite continuum). Indeed this is where the methods discussed in Drummond et al. (2006), where rates may change but only at branching events, can be seen as somewhat of an intelligent compromise between a (statistically tractable) homogeneous model and a (biologically realistic) inhomogeneous model. Another approach would be to abandon the continuous-time hypothesis and work with discrete Markov chains (or equivalently "algebraic" models, Pachter and Sturmfels, 2005). However this approach introduces many free parameters and suffers from a lack of interpretation, as it is unclear what the free algebraic parameters mean in biological terms (such as divergence times and molecular rates), except with reference to the corresponding continuous-time approach.

An available resolution of these issues is to observe that it is possible to continue to model phylogenetic processes as being homogeneous, but interpret the transition rates that are fitted on each of the tree as a kind of "average" of the true inhomogeneous process for that edge. It is this perspective that we take in this work and it leads directly to the concept of *multiplicative closure* for continuous-time Markov chains. It will be shown that models that are multiplicatively closed have the property that, even in their inhomogeneous formulation, it is possible to interpret their average behavior as a homogeneous process. It is then the purpose of this paper to discuss sufficient conditions for multiplicative closure of continuous-time Markov chains. In order to generate particular examples of closed models, we exploit symmetry properties of DNA substitution rates to present a scheme that creates a hierarchy of closed Markov models based on the number of free parameters available. We also note that our results are applicable to any application of continuous-time Markov chains, and as such are of general relevance beyond the phylogenetic applications that we discuss to motivate this study.

In Section 2 we give basic definitions of multiplicative closure and Lie Markov models. To achieve this we review the required Lie theory, and we discuss the Lie algebra of the general Markov model. As an example to motivate the general procedure, in Section 3 we specialize to the case of binary Markov chains and give a complete description of Lie Markov models in this case. In Section 4 we discuss the symmetry properties of DNA models under nucleotide permutations, and show that these symmetries are statistically relevant to likelihood calculations. We then explain how such symmetry can be used to assist in the search for Lie Markov models. Here we also prove that "equivariant" (see Draisma and Kuttler, 2008; Casanellas and Fernández-Sánchez, 2010) and "group-based" (see Semple and Steel, 2003) models are examples of Lie Markov models. In Section 5 we give a general scheme for generating a full list of Lie Markov models with a given symmetry property. In Section 6 we explicitly give four state Lie Markov models with maximal symmetry. Finally, Section 7 discusses implications and possibilities for future work.

2. Lie algebras and closure of Markov models

For algebraic simplicity we work over the complex field $\mathbb{C}$, and refer to a matrix as "Markov" if it has unit column sums. Later we will discuss how our discussion specializes to the stochastic case where the entries must be real and lie in the range [0,1]. Rather than work directly with the general Markov model, we will also consider only Markov matrices that have non-zero determinant. Although this need not be the case for a general Markov matrix, it is not too stringent a condition as (i) the set of Markov matrices with zero determinant is of measure zero in the set of Markov matrices (this is because they are defined by the vanishing of a single polynomial function and hence lie in an ambient space of dimension one less than the set of generic Markov matrices), (ii) Markov matrices that arise from a continuous-time formulation have non-zero determinant (as we will see shortly). In any case, in the conclusions we will argue that understanding Markov matrices with zero determinant becomes easier once we understand how the rest can be categorized.

Let the *general Markov model* $\mathfrak{M}_{GMM}$ be the set of $n \times n$ matrices with column sum 1

$$\mathfrak{M}_{GMM} \coloneqq \{M \in \mathbb{M}_n(\mathbb{C}) : \boldsymbol{\theta}^T M = \boldsymbol{\theta}^T\},$$

where $\boldsymbol{\theta}$ is the column n-vector with all its entries equal to 1, i.e. $\boldsymbol{\theta}^T = (1, 1, \ldots, 1)$. Specializing further, consider the subset of matrices in $\mathfrak{M}_{GMM}$ with non-zero determinant

$$GL_1(n,\mathbb{C}) \coloneqq \{M \in \mathbb{M}_n(\mathbb{C}) : \boldsymbol{\theta}^T M = \boldsymbol{\theta}^T, \det(M) \neq 0\}.$$

In turn, this set of matrices includes a subset of matrices that arise by taking the exponential of a rate-matrix; that is, the exponential of a matrix in

$$\mathfrak{Q}_{GMM} \coloneqq \{Q \in M_n(\mathbb{C}) : \boldsymbol{\theta}^T Q = \mathbf{0}^T\}. \tag{1}$$

We will refer to $e^{\mathfrak{Q}_{GMM}} \coloneqq \{e^Q : Q \in \mathfrak{Q}_{GMM}\}$ as "the general rate-matrix model" and below we will discuss matrix exponentials in more detail (particularly their importance to Lie theory).

As the inverse of a Markov matrix (if it exists) is also a Markov matrix it is clear that $GL_1(n,\mathbb{C})$ is actually a subgroup of the general linear group $GL(n,\mathbb{C})$, and it follows that $GL_1(n,\mathbb{C})$ and $e^{\mathfrak{Q}_{GMM}}$ are actually *Lie groups* (see Stillwell, 2008, for the relevant technical definitions). In fact we have the isomorphism $GL_1(n,\mathbb{C}) \cong A(n\text{–}1,\mathbb{C})$ where $A(n\text{–}1,\mathbb{C})$ is the (complex) *affine group* (see for example, Baker, 2003). This observation allows the general methods of Lie theory to be applied to understanding continuous-time Markov models; see Johnson (1985) and Mourad (2004) for general results and discussion, and Sumner et al. (2008) and Sumner and Jarvis (2009) for applications to phylogenetics.

Summarizing, we have the following set inclusions:

$$e^{\mathfrak{Q}_{GMM}} \subset GL_1(n,\mathbb{C}) \subset \mathfrak{M}_{GMM},$$

and Lie group hierarchy

$$e^{\mathfrak{Q}_{GMM}} < GL_1(n,\mathbb{C}) < GL(n,\mathbb{C}).$$

We define a *Markov model* $\mathfrak{M}$ by taking $\mathfrak{M} \subseteq \mathfrak{M}_{GMM}$ as some well defined subset of the general Markov model. Similarly, a *rate-matrix model* $e^{\mathfrak{Q}}$ is defined by taking $\mathfrak{Q} \subseteq \mathfrak{Q}_{GMM}$ as some well defined subset of rate-matrices drawn from the general rate-matrix model and taking the set of exponentials thereof (as in (1)). It follows immediately from these definitions that all rate-matrix models are Markov models. In what follows we are primarily interested in the case that $\mathfrak{M} = e^{\mathfrak{Q}}$, and in this case we will abuse our terminology and refer to $\mathfrak{Q}$ as a "model".

Definition 2.1. A Markov model $\mathfrak{M}$ is said to be *multiplicatively closed* if and only if for all $M_1, M_2 \in \mathfrak{M}$ we also have $M_1 M_2 \in \mathfrak{M}$.

Of course, recalling that matrix multiplication is associative, this is exactly the statement that $\mathfrak{M}$ forms a *semigroup* under matrix multiplication, and is identical to the definition of "substitution model" given in Gronau et al. (2009).