GENERATING MARKOV EVOLUTIONARY MATRICES FOR A GIVEN BRANCH LENGTH

MARTA CASANELLAS AND ANNA KEDZIERSKA

Abstract. Under a markovian evolutionary process, the expected number of substitutions per site (branch length) that occur when a sequence evolves from another via a transition matrix P can be approximated by $-1/4 \log(\det P)$. In continuous-time models, it is easy to simulate the process for any given branch length. For discrete-time models, it is not so trivial. In this paper we solve this problem for the most well-known discrete-time models `JC69`*, `K80`*, `K81`*, `SSM`, and `GMM` and we provide concise algorithms to generate stochastic matrices of given determinant. These models have the advantage to be nonhomogeneous evolutionary processes.

Keywords: stochastic matrix, Markov matrix, branch length, determinant, evolutionary model.
AMS subj. class.: 15B51, 60J10.

1. Introduction

Phylogenetic reconstruction methods are usually tested on simulated data, i.e. DNA (or protein) sequences that have been randomly generated following a molecular evolutionary model on a phylogenetic tree. It is easy to generate a random DNA sequence that evolves from a given DNA sequence under a particular Markov evolutionary model if no more constrains are required: one just needs to give random values to the parameters of the model and generate data according to the conditional probabilities obtained from the parameters. An extra effort is needed if the amount of "substitution events" is fixed; this magnitude is usually called the *branch length* of the edge relating both sequences in the phylogenetic tree.

We assume (as it is commonly done) that sites in a DNA sequence are independent and identically distributed (*iid* hypothesis), so that one just models the evolution of one site (thought as a random variable taking values in $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$). The most common molecular evolutionary models used in phylogenetics are the so-called *continuous-time models*. In these models, the substitution events along an edge e of a rooted phylogenetic tree occur following a continuous-time Markov

process: there is an instantaneous mutation rate matrix Q (usually fixed throughout the tree) that operates at intensity λ_e and for duration t_e so that the substitution matrix (or transition matrix) P_e equals $\exp(Q \cdot \lambda_e t_e)$. Among them there are the (stationary and time-reversible) models Jukes-Cantor `JC69` [10], Kimura two-parameters `K80` [13], Kimura three-parameters `K81` [14], `HKY` [9], and `GTR` [20].

In this paper we consider a broader class of evolutionary models, the *(discrete-time)* Markov models on phylogenetic trees. These models have been widely used in phylogenetics, but mostly in the theoretical side (see [17, chapter 8] for an overview of Markov processes on trees). Briefly, the parameters of these models consist of a rooted tree topology, a root distribution, and substitution matrices P_e on the edges e of the tree whose entries correspond to the conditional probabilities $P(x|y, e)$ that a nucleotide y at the parent node of e is substituted by nucleotide x at the child node. In particular, there is no instantaneous rate matrix fixed on the whole tree for these models, so that they account for what is called *nonhomogeneous* data: different lineages in the tree are allowed to evolve at different rates. Moreover, there is no stationary distribution implicitly assumed. These are some of the main advantages of these models as opposite to the most used continuous-time models. We refer to [2] and [17, chapter 8] for a mathematical approach to the evolutionary models used in this paper.

If a DNA sequence has evolved from another according to a substitution matrix P_e, then the number of substitutions per site that have occurred can be approximated by

$$l(e) = -\frac{1}{4} \log \det(P_e) \tag{1}$$

(see [3]). This is usually known as the *branch length* of edge e measured in the expected number of substitutions per site. In the case of stationary and time-reversible continuous-time models, the expected number of substitutions coincides with $-\operatorname{tr}(D(\Pi)Q\lambda_e t_e)$ if $P_e = \exp(Q \cdot \lambda_e t_e)$ and $D(\Pi)$ is a diagonal matrix with entries corresponding to the stationary distribution Π (see [17, chapter 8]).

Generating DNA sequences evolving under a stationary and time-reversible continuous-time model on an edge e with preassigned branch length l and rate matrix Q, is not difficult: one just needs to take $\lambda_e t_e = -l/\operatorname{tr}(D(\Pi)Q)$ and follow the usual process to generate a Poisson distribution according to these parameters. There are several programs available for generating data under most-used continuous-time evolutionary models, for example `seq-gen` [16] and `evolver` in *PAML* [21].

Here we deal with the problem of generating data evolving under the more general discrete-time models when the branch lengths of the tree are fixed. From

what we have seen above, this problem is equivalent to the generation of substitution matrices P_e (belonging to the evolutionary model) with given determinant. As the substitution matrices are stochastic matrices (nonnegative entries and rows summing to one), this is not an easy task. We solve this problem for the so-called equivariant models JC69*, K81*, K80* and SSM ([8],[5]), and for the general Markov model GMM ([3], [18], [1]). Models JC69*, K81*, K80* correspond to the discrete-time version of the corresponding continuous-time models, and SSM contains HKY as a submodel. Our results for the first four models (Propositions 3.1, 4.2, 5.1, and 6.6) are actually bidirectional: we provide algorithms for generating *any* strictly stochastic matrix M with determinant equal to a given number $K \in (0,1)$, when M is either a JC69*, K81*, K80* or SSM matrix. For the most general model GMM we provide a way of generating strictly stochastic matrices with determinant equal to K, but we are not able to claim whether we produce all of them. We observe that we are able to produce matrices that are not the exponential of a real rate matrix (cf. Remark 5.5).

The ingredients for the proofs in this paper rely on invariant subspace decomposition techniques for the SSM and diagonalization for JC69*, K81* and K80* models. In section 2 we provide the background on the evolutionary models considered in this paper and we recall the diagonal form of JC69*, K81*, K80* models. Our results for JC69*, K81*, K80*, SSM, and GMM models are given in sections 3 to 7 respectively.

The algorithms proposed in this paper have been implemented in C++ in order to generate multiple sequence alignments of DNA data evolving on any phylogenetic tree, see http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgGenNonH.pl. In this software we have addressed the problem of producing biologically meaningful matrices, namely rerunning the algorithm until a matrix whose diagonal entries are the largest in each column is produced (see [11]). The quoted software has been already used to simulate data for testing model selection methods in [12].

2. Preliminaries

Definition 2.1. A 4×4 matrix A with real entries and row sums equal to 1,

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{pmatrix},$$

$\sum_j a_{i,j} = 1$, is called a GMM matrix. The GMM matrix above is called a SSM matrix if $a_{3,1} = a_{2,4}$, $a_{3,2} = a_{2,3}$, $a_{3,3} = a_{2,2}$, $a_{3,4} = a_{2,1}$, $a_{4,1} = a_{1,4}$, $a_{4,2} = a_{1,3}$, $a_{4,3} = a_{1,2}$,

$a_{4,4} = a_{1,1}$. If moreover $a_{1,1} = a_{2,2}$, $a_{1,2} = a_{2,1}$, $a_{1,3} = a_{2,4}$ and $a_{1,4} = a_{2,3}$, then A is called a K81* matrix. If a K81* matrix satisfies $a_{1,2} = a_{1,4}$, then it is called a K80* matrix and it is called a JC69* matrix if also $a_{1,2} = a_{1,3}$.

In other words, an SSM matrix is a matrix of type

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix} \quad \text{with} \quad \begin{array}{l} a+b+c+d=1 \\ e+f+g+h=1 \end{array};$$

a K81* matrix is a matrix of type

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix} \quad \text{with } a+b+c+d=1;$$

a K80* matrix is a matrix of type

$$\begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix} \quad \text{with } a+2b+c=1;$$

and a JC69* matrix is a matrix of type

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \quad \text{with } a+3b=1.$$

The names of the matrices above come from well known evolutionary models: in the stochastic case, GMM is a transition matrix for the general Markov model ([3], [18], [1]), SSM for the strand symmetric model introduced in [6], K81* for the discrete-time version of Kimura three-parameters model [14], K80* for the discrete-time version of Kimura two-parameters model [13], and JC69* for the discrete-time version of Jukes-Cantor model [10].

Definition 2.2. A square matrix A is called a *stochastic matrix* if it has row sums equal to 1 and nonnegative real entries. It is called *strictly stochastic* if moreover all its entries are strictly positive.

We recall that the determinant of any stochastic matrix has absolute value less than or equal to 1 (this is a consequence of Perron-Frobenius theorem). In this paper we address the problem of providing stochastic matrices of the above shapes with given determinant $K \in (0, 1)$.

Before ending the preliminaries section we want to point out in the lemma below that JC69*, K80* and K81* matrices are diagonalizable.

Lemma 2.3. *Let* $A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$ *be a* K81* *matrix* $(a+b+c+d=1)$ *and consider the matrix*

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

Then $S^{-1} = \frac{1}{4}S$ *and* $S^{-1}AS$ *is a diagonal matrix with diagonal entries* $\{1, a-b-c+d, a-b+c-d, a+b-c-d\}$ *(in this order).*

Remark 2.4. The change of variables considered in the Proposition above corresponds to the discrete Fourier transform in the setting of [19].

3. Generating JC69* matrices with given determinant

Proposition 3.1. *Let* $K \in (0,1)$ *and let*

$$A = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}, \quad a+3b=1,$$

be a JC69* *matrix. Then* A *is a strictly stochastic matrix with determinant equal to* K *if and only if* $a = \frac{1}{4}(1+3K^{1/3})$, $b = \frac{1-a}{3}$.

Proof. Using Lemma 2.3 we have $\det A = (\frac{4a-1}{3})^3$. Therefore, A has determinant equal to K if and only if $a = \frac{1}{4}(1+3K^{1/3})$. Moreover, as $K \in (0,1)$, we obtain $1 > a > 0$ (and so $0 < b = \frac{1-a}{3} < 1$), and we are done. □

In this case, all stochastic JC69* substitution matrices are of exponential type. According to the result above, it is easy to generate stochastic JC69* matrices with given determinant:

Algorithm 3.2. (Generation of JC69* matrices with given determinant.)

Input: K in $(0,1)$.

Output: A strictly stochastic JC69* matrix A with determinant K.

Step 1: Set $a = \frac{1}{4}(1+3K^{1/3})$, $b = \frac{1-a}{3}$.

`Final`: Return

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

4. Generating K80* matrices with given determinant

Remark 4.1. As a technical step previous to the generation of K80* matrices with given determinant, we consider the polynomial

$$p_K(x) = -2x^3 + x^2 + K, \quad K \in (0,1),$$

and we observe that it has exactly one real root s which lies in $(\sqrt{K}, 1)$. Indeed, the coefficients of $p_K(x)$ have one variation in sign and those of $p_K(-x)$ have no variation in sign. Therefore, applying Descartes' rule we obtain that $p_K(x)$ has exactly one positive root s and no negative roots. Moreover, as K is a constant in $(0,1)$, we have that $p_K(\sqrt{K}) = 2K(1-\sqrt{K})$ is positive and $p_K(1) = K - 1$ is negative, implying that s lies in $(\sqrt{K}, 1)$.

Using the formula for the roots of a cubic polynomial we obtain

$$s = \frac{1}{6} + \frac{1}{6}\sqrt[3]{1 + 54K + 6\sqrt{3K + 81K^2}} + \frac{1}{6}\sqrt[3]{1 + 54K - 6\sqrt{3K + 81K^2}}.$$

As a byproduct, the polynomial $p_K(-x)$ has exactly one real root which coincides with $-s$.

Proposition 4.2. *Let $K \in (0,1)$ and let s be the unique real root of $p_K(x) = -2x^3 + x^2 + K$ (see Remark 4.1). Let*

$$A = \begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix},$$

be a K80 matrix ($a + 2b + c = 1$), and consider the change of variables $\alpha = 1 - 2(b + c)$, $\beta = 1 - 4b$. Then A is a strictly stochastic matrix with determinant equal to K if and only if $\sqrt{K} < |\alpha| < s$ and $\beta = K/\alpha^2$.*

Remark 4.3. Note that by Lemma 2.3, the matrix A above diagonalizes with eigenvalues 1, α (with multiplicity 2) and β. Therefore in this case the matrix A has a real logarithm. Indeed, according to [7], a non-singular diagonalizable matrix is the exponential of a real matrix if and only if its negative eigenvalues occur with even multiplicity (above, this is the case when $\alpha < 0$).

Proof. First we note that the inverse change of variables is $b = \frac{1-\beta}{4}$, $c = \frac{1+\beta-2\alpha}{4}$. Moreover, $\alpha = a - c$ and $\beta = a - 2b + c$ are the diagonal entries in $S^{-1}AS$ (different than 1) in Lemma 2.3 and therefore $\det(A) = \alpha^2\beta$.

$\Rightarrow$) Assume that A is strictly stochastic with determinant K. Then b is strictly positive, so that $\beta < 1$. As $K = \det(A) = \alpha^2\beta$ and $\beta < 1$, we obtain $|\alpha| > \sqrt{K}$. In particular, $\alpha \neq 0$ and we can write $\beta = K/\alpha^2$.

Using the inverse change of variables above and $\beta = K/\alpha^2$ we have

$$a > 0 \Leftrightarrow 2b + c < 1 \Leftrightarrow \frac{3 - K/\alpha^2 - 2\alpha}{4} < 1 \Leftrightarrow p_K(-\alpha) > 0.$$

As noted in Remark 4.1, $p_K(-x)$ has exactly one negative root which equals $-s$ and lies in $(-1, -\sqrt{K})$. As $p_K(-x)$ has positive leading term, $p_K(-\alpha) > 0$ only holds if $\alpha > -s$.

Similarly, c is strictly positive if and only if $p_K(\alpha) > 0$. Following an analogous argument, we obtain that $p_K(\alpha) > 0$ if and only if $\alpha < s$. Putting all together we obtain $\sqrt{K} < |\alpha| < s$, as desired.

$\Leftarrow$) Assume that $\sqrt{K} < |\alpha| < s$ and $\beta = K/\alpha^2$. In particular, we have $\beta < \frac{K}{K} = 1$ and we obtain that $b = \frac{1-\beta}{4}$ is strictly positive.

Now, as in the proof of $\Rightarrow$) we have that $c > 0$ if and only if $p_K(\alpha) > 0$. And also as above, this happens if and only if $\alpha < s$. As we assumed $|\alpha| < s$, we obtain $c > 0$.

Lastly, $a > 0$ if and only of $p_K(-\alpha) > 0$, and this holds if and only if $\alpha > -s$ (see proof of $\Rightarrow$). As we assumed $|\alpha| < s$, we get that A is a strictly stochastic matrix.

Moreover, $\det(A) = \alpha^2\beta = K$ as wanted. □

Using the previous result, we provide the following algorithm for generating strictly stochastic $\texttt{K80}^*$ matrices with given determinant K. It is worth pointing out that with this algorithm we are generating *all* $\texttt{K80}^*$ strictly stochastic matrices with determinant K.

Algorithm 4.4. (Generation of $\texttt{K80}^*$ matrices with given determinant.)

Input: K in $(0, 1)$.

Output: A strictly stochastic $\texttt{K80}^*$ matrix A with determinant K.

`Step 1:` Compute the unique real root s of $p_K(x)$ using Remark 4.1.
`Step 2:` Choose α randomly such that $\sqrt{K} < |\alpha| < s$.

`Step 3:` Let $\beta := K/\alpha^2$, $b := \frac{1-\beta}{4}$, $c := \frac{1+\beta-2\alpha}{4}$, and $a := 1-2b-c$.
`Final:` Return

$$A := \begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix}.$$

5. Generating K81* matrices with given determinant

Previously to dealing with the case of K81* matrices, for each real number K in $(0,1)$, we let s be the unique positive root of the polynomial

$$q_K(z) := z(z+1)^2 - 4K.$$

Indeed, according to Descartes' rules of signs, this polynomial has at most one positive root. Moreover, as $q_K(K) < 0$ and $q_K(1) > 0$, there is exactly one positive root s and it lies in $(K,1)$. Using the formula for the roots of a cubic polynomial we obtain

$$s = -\frac{2}{3} - \frac{1}{3}\sqrt[3]{-1-54K+6\sqrt{3K+81K^2}} - \frac{1}{3}\sqrt[3]{-1-54K-6\sqrt{3K+81K^2}}. \tag{2}$$

Proposition 5.1. *Let $K \in (0,1)$ and let s be the unique real root of $q_K(z) := z(z+1)^2 - 4K$. Let*

$$A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix},$$

be a K81 matrix ($a+b+c+d=1$), and consider the change of variables $\alpha = 1-2(b+c)$, $\beta = 1-2(b+d)$, $\gamma = 1-2(c+d)$. Then A is a strictly stochastic matrix with determinant equal to K if and only if $|\alpha| \in (s,1)$, $|\beta| \in \left(I_{|\alpha|}, J_{|\alpha|}\right)$ where*

$$I_{|\alpha|} = \max\left\{ \frac{-1+|\alpha|+\sqrt{(1-|\alpha|)^2+\frac{4K}{|\alpha|}}}{2}, \frac{1+|\alpha|-\sqrt{(1+|\alpha|)^2-\frac{4K}{|\alpha|}}}{2} \right\},$$

$$J_{|\alpha|} = \min\left\{ \frac{1+|\alpha|+\sqrt{(1+|\alpha|)^2-\frac{4K}{|\alpha|}}}{2}, \frac{1-|\alpha|+\sqrt{(1-|\alpha|)^2+\frac{4K}{|\alpha|}}}{2} \right\},$$

and $\gamma = \frac{K}{\alpha\beta}$.

Remark 5.2. As the change of variables above is symmetric in b, c, d, the roles of these three variables can be exchanged in the previous Proposition.

Before proving this Proposition we need the following technical lemma.

Lemma 5.3. *Let K be a real number in $(0,1)$, let s be the unique positive solution to $z(z+1)^2 - 4K = 0$, and consider the function*

$$f(x,y) = 1 - x - y + \frac{K}{xy}$$

defined over $\mathbb{R}^2_+$.Given $y \in (0,1)$, we consider the set

$$\Omega_y = \{x \in \mathbb{R} \mid x > 0, f(x,y) > 0, f(x,-y) > 0, f(-x,y) > 0, f(-x,-y) > 0\}.$$

Then Ω_y is not empty if and only if $y > s$. Moreover, if $x \in \Omega_y$, then x belongs to (I_y, J_y) where

$$I_y = \max\left\{\frac{-1+y+\sqrt{(1-y)^2+\frac{4K}{y}}}{2}, \frac{1+y-\sqrt{(1-y)^2-\frac{4K}{y}}}{2}\right\} \quad \textit{and}$$

$$J_y = \min\left\{\frac{1+y+\sqrt{(1-y)^2-\frac{4K}{y}}}{2}, \frac{1-y+\sqrt{(1-y)^2+\frac{4K}{y}}}{2}\right\}.$$

Proof. We fix $y > 0$ and we view f as a function on x. We define the quadratic functions $\tilde{f}_y(x) := -x^2 + (1-y)x + K/y$ and $\tilde{g}_y(x) := x^2 + (1+y)x + K/y$ so that $\tilde{f}_y(x) = xf(x,y)$ and $\tilde{g}_y(x) = xf(-x,-y)$. In particular, x belongs to Ω_y if and only if $x > 0$, $\tilde{f}_y(x) > 0$, $\tilde{f}_{-y}(x) > 0$, $\tilde{g}_y(x) > 0$ and $\tilde{g}_{-y}(x) > 0$.

Note that $\tilde{f}_y$ has discriminant $\Delta_1(y) = (1-y)^2 + \frac{4K}{y}$ and $\tilde{g}_y$ has discriminant $\Delta_2(y) = (1+y)^2 - \frac{4K}{y}$.

We observe that $\Delta_1(y) > 0$ for $y > 0$. Therefore $\tilde{f}_y(x) = 0$ has the real solutions $x_{1,L}(y) = \frac{1-y-\sqrt{\Delta_1(y)}}{2}$, $x_{1,R}(y) = \frac{1-y+\sqrt{\Delta_1(y)}}{2}$, and $\tilde{f}_y(x)$ is positive for x in $(x_{1,L}, x_{1,R})$. Note that $\sqrt{\Delta_1(y)} > |1-y|$ for $y > 0$, so $x_{1,L}(y)$ is negative and $x_{1,R}(y)$ is positive. Therefore, for $x > 0$ and $y > 0$, $\tilde{f}_y(x)$ is positive if and only if $x \in (0, x_{1,R}(y))$.

On the other hand, as $\tilde{f}_{-y}$ has negative leading coefficient, there exists x with $\tilde{f}_{-y}(x) > 0$ if and only if $\Delta_1(-y) > 0$. Note that $\Delta_1(-y)$ is positive for $y > 0$ if and only if $y > s$ (indeed, $\Delta_1(-y)$ coincides with $q_K(y)/y$).

Thus $\tilde{f}_{-y}(x) > 0$ has a solution for $x > 0$, if and only if $y > s$. Now for $x > 0, y > s$, the roots of $\tilde{f}_{-y}(x) = 0$ are $x_{1,L}(-y)$ and $x_{1,R}(-y)$. Clearly $x_{1,R}(-y)$ and $x_{1,L}(-y)$ are both positive for $y > s$. Therefore, for $x > 0$ and $y > 0$, we have $\tilde{f}_{-y}(x) > 0$ if and only if $y > s$ and $x \in (x_{1,L}(-y), x_{1,R}(-y))$.

Now we study the positivity of $\tilde{g}_y(x)$ for $x > 0$. Note that $\tilde{g}_y$ has discriminant $\Delta_1(-y)$. As the leading coefficient of $\tilde{g}_y$ is positive, we have that $\tilde{g}_y(x) > 0$ for all $y < s$ and $x \in \mathbb{R}$ (because in this case the discriminant is negative). Moreover, if $y > s$, the real roots of $\tilde{g}_y(x) = 0$ are $x_{2,L}(y) = \frac{-(1+y)-\sqrt{\Delta_1(-y)}}{2}$ and $x_{2,R}(y) = \frac{-(1+y)+\sqrt{\Delta_1(-y)}}{2}$. They are both negative so that $\tilde{g}_y(-x)$ is positive for all $y > s$ and $x > 0$.

We study the positivity of $\tilde{g}_{-y}(x)$ for $x > 0$ and $y > 0$. The discriminant of $\tilde{g}_{-y}$ is $\Delta_1(y)$, and it is positive for $y > 0$. Then the roots of $\tilde{g}_{-y}$ are $x_{2,L}(-y)$ and $x_{2,R}(-y)$. For $y > 0$ we have $x_{2,L}(-y) < 0$ and $x_{2,R}(-y) > 0$, and therefore $\tilde{g}_{-y}(x) > 0$ if and only if x belongs to $(x_{2,R}(-y), +\infty)$.

Summing up, we have proven that the set Ω_y is non-empty if and only if $y > s$. Moreover, in that case, if x belongs to Ω_y, then x lies in

$$(0, x_{1,R}(y)) \cap (x_{1,L}(-y), x_{1,R}(-y)) \cap (0, +\infty) \cap (x_{2,R}(-y), +\infty)\,.$$

It is easy to see that $x_{1,R}(y)$ is bigger than $x_{2,R}(-y)$ for $y \in (0, 1)$. Therefore the intersection of intervals above is equal to

$$(x_{1,L}(-y), x_{1,R}(-y)) \cap (x_{2,R}(-y), x_{1,R}(y))\,.$$

It is easy to check that, for $y \in (0, 1)$, one has $x_{2,R}(-y) < x_{1,R}(-y)$ and $x_{1,L}(-y) < x_{1,R}(-y)$, which ends the proof of the lemma. □

Proof of Proposition 5.1. Taking into account that $a = 1 - b - c - d$, we note that inverse change of variables is $a = \frac{1}{4}(1 + \alpha + \beta + \gamma)$, $b = \frac{1}{4}(1 - \alpha - \beta + \gamma)$, $c = \frac{1}{4}(1 - \alpha + \beta - \gamma)$, $d = \frac{1}{4}(1 + \alpha - \beta - \gamma)$. Observing that α, β, γ are the diagonal entries in $S^{-1}AS$ in Lemma 2.3, we see that $\det A = \alpha\beta\gamma$.

⇒) Assume that A is stochastic with determinant $K \in (0, 1)$. Then α, β, and γ are non-zero, and $\gamma = \frac{K}{\alpha\beta}$. From the positivity of a, b, c, d we get that $1 + \alpha + \beta + \frac{K}{\alpha\beta} > 0$, $1 - \alpha - \beta + \frac{K}{\alpha\beta} > 0$, $1 - \alpha + \beta - \frac{K}{\alpha\beta} > 0$, and $1 + \alpha - \beta - \frac{K}{\alpha\beta} > 0$. In terms of Lemma 5.3, these inequalities can be rewritten as

$$f(-\beta, -\alpha) > 0, f(\beta, \alpha) > 0, f(\beta, -\alpha) > 0, f(-\beta, \alpha) > 0.$$

Therefore $|\beta|$ is an element of $\Omega_{|\alpha|}$, which implies that $|\alpha| > s$ (see Lemma 5.3). Moreover, as $\alpha = 1 - 2(b + d)$, and $b, d > 0$, we see that $|\alpha| < 1$. The result then follows from Lemma 5.3.

⇐) Using Lemma 5.3 we see that under these assumptions, $\Omega_{|\alpha|} \neq \emptyset$ and $|\beta|$ belongs to $\Omega_{|\alpha|}$. Therefore $f(-\beta, -\alpha) > 0, f(\beta, \alpha) > 0, f(\beta, -\alpha) > 0, f(-\beta, \alpha) >$

0. As $\gamma = \frac{K}{\alpha\beta}$, these inequalities coincide with $a > 0$, $b > 0$, $c > 0$ and $d > 0$, and we are done. □

The previous results give us a way of generating *any* stochastic K81* matrix.

Algorithm 5.4. (Generation of K81* matrices with given determinant.)

Input: K in $(0, 1)$.

Output: A strictly stochastic K81* matrix A with determinant K.

Step 1: Compute the unique real root s of $z(z+1)^2 - 4K$ using (2).
Step 2: Choose α randomly such that $1 > |\alpha| > s$.
Step 3: Take β randomly such that $|\beta|$ belongs to $(I_{|\alpha|}, J_{|\alpha|})$.
Step 4: Set $\gamma = \frac{K}{\alpha\beta}$.
Step 5: Set $a = \frac{1}{4}(1 + \alpha + \beta + \gamma)$, $b = \frac{1}{4}(1 - \alpha - \beta + \gamma)$, $c = \frac{1}{4}(1 - \alpha + \beta - \gamma)$, $d = \frac{1}{4}(1 + \alpha - \beta - \gamma)$.
Final: Return

$$A = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}.$$

Remark 5.5. The change of variables in Proposition 5.1 diagonalizes the matrix to $\mathrm{Diag}(1, \alpha, \beta, \gamma)$ (see Lemma 2.3). As we have seen in that proposition, α and β can be both negative. Therefore, using the same arguments as in Remark 4.3, we observe that the matrices produced by the algorithm above are not all of type $\exp(Q)$ for a real matrix Q.

6. Generating SSM matrices with given determinant

Definition 6.1. Let A be a 4×4 real matrix. We call $F(A)$ the matrix obtained from A after performing the basis change $F(A) = S^{-1}AS$ where

$$S = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

When A is an SSM matrix, A can be viewed as an element in $Hom_G(\mathbb{C}^4, \mathbb{C}^4)$ where $G =< (\mathtt{AT})(\mathtt{CG}) >$ (see [4]). The change of basis above decomposes $\mathbb{C}^4$ into its isotypic components via the natural linear representation $G \longrightarrow GL(\mathbb{C}^4)$. This change of basis is also known as the generalized Fourier transform (see [6]). We have the following fact:

Lemma 6.2. *A 4×4 matrix $A=(a_{i,j})$ is an* SSM *matrix if and only if $F(A)$ has the following shape:*

$$F(A)=\begin{pmatrix} \lambda & 1-\lambda & 0 & 0\\ 1-\mu & \mu & 0 & 0\\ 0 & 0 & \alpha & \alpha'\\ 0 & 0 & \beta' & \beta \end{pmatrix}.$$

In this case, λ,μ, α,α', β,β' can be written in terms of the entries of A as $\lambda=a_{1,1}+a_{1,4}$, $\mu=a_{2,2}+a_{2,3}$, $\alpha=a_{2,2}-a_{2,3}$, $\alpha'=a_{2,4}-a_{2,1}$, $\beta=a_{1,1}-a_{1,4}$, and $\beta'=a_{1,3}-a_{1,2}$. The inverse change of variables is $a_{1,1}=(\lambda+\beta)/2$, $a_{1,2}=(1-\lambda-\beta')/2$, $a_{1,3}=(1-\lambda+\beta')/2$ $a_{1,4}=(\lambda-\beta)/2$, $a_{2,1}=(1-\mu-\alpha')/2$, $a_{2,2}=(\mu+\alpha)/2$, $a_{2,3}=(\mu-\alpha)/2$, $a_{2,4}=(1-\mu+\alpha')/2$. Moreover, A is a strictly stochastic SSM *matrix if and only if $\lambda,\mu\in(0,1)$, $|\beta|<\lambda$, $|\beta'|<1-\lambda$, $|\alpha|<\mu$, and $|\alpha'|<1-\mu$.*

Proof. The matrix $F(A)$ for a generic matrix $A=(a_{i,j})$ is

$$\frac{1}{2}\left(\begin{array}{cc|cc} a_{1,1}+a_{1,4}+a_{4,1}+a_{4,4} & a_{1,2}+a_{1,3}+a_{4,2}+a_{4,3} & a_{1,2}-a_{1,3}+a_{4,2}-a_{4,3} & a_{1,4}-a_{1,1}-a_{4,1}+a_{4,4}\\ a_{2,1}+a_{2,4}+a_{3,1}+a_{3,4} & a_{2,2}+a_{2,3}+a_{3,2}+a_{3,3} & a_{2,2}-a_{2,3}+a_{3,2}-a_{3,3} & a_{2,4}-a_{2,1}-a_{3,1}+a_{3,4}\\ \hline a_{2,1}+a_{2,4}-a_{3,1}-a_{3,4} & a_{2,2}+a_{2,3}-a_{3,2}-a_{3,3} & a_{2,2}-a_{2,3}-a_{3,2}+a_{3,3} & a_{2,4}-a_{2,1}+a_{3,1}-a_{3,4}\\ a_{4,1}+a_{4,4}-a_{1,1}-a_{1,4} & a_{4,2}+a_{4,3}-a_{1,2}-a_{1,3} & a_{1,3}-a_{1,2}+a_{4,2}-a_{4,3} & a_{1,1}-a_{1,4}-a_{4,1}+a_{4,4} \end{array}\right).$$

If A is an SSM matrix, then $a_{3,1}=a_{2,4}$, $a_{3,2}=a_{2,3}$, $a_{3,3}=a_{2,2}$, $a_{3,4}=a_{2,1}$, $a_{4,1}=a_{1,4}$, $a_{4,2}=a_{1,3}$, $a_{4,3}=a_{1,2}$, and $a_{4,4}=a_{1,1}$. Therefore the non-diagonal blocks are 0. Moreover, as sums of rows are equal to 1, we have that the entries of each row in the upper left block sum to 1:

$$\frac{1}{2}(a_{1,1}+a_{1,4}+a_{4,1}+a_{4,4}+a_{1,2}+a_{1,3}+a_{4,2}+a_{4,3})=1,$$

$$\frac{1}{2}(a_{2,1}+a_{2,4}+a_{3,1}+a_{3,4}+a_{2,2}+a_{2,3}+a_{3,2}+a_{3,3})=1.$$

Conversely, imposing that the entries of non-diagonal blocks in $F(A)$ are equal to 0 is equivalent to imposing $a_{3,1}=a_{2,4}$, $a_{3,2}=a_{2,3}$, $a_{3,3}=a_{2,2}$, $a_{3,4}=a_{2,1}$, $a_{4,1}=a_{1,4}$, $a_{4,2}=a_{1,3}$, $a_{4,3}=a_{1,2}$, and $a_{4,4}=a_{1,1}$ (adding and subtracting certain pairs of equations). Moreover, $F(A)_{1,1}+F(A)_{1,2}=1$ implies that sum of rows 1 and 4 is equal to 2 (and similar for rows 2 and 3). But we have just seen that the set of entries in the first (resp. second) row is equal to the set of entries in the forth (resp. third) row, thus the sum of entries in each row is equal to 1.

Finally, if A is strictly stochastic,then we clearly have λ and μ in $(0,1)$, $|\alpha|<\mu$, $|\alpha'|<1-\mu$, $|\beta|<\lambda$, and $|\beta'|<1-\lambda$. Conversely, if these inequalities are satisfied for $F(A)$, using the inverse change of variables, we see that A is strictly stochastic.

□

Before stating the main result of this section we introduce some notation and we prove a technical result.

Remark 6.3. Given $K \in (0,1)$, we consider the polynomial $r_K(z) = z^3 + z - 2K$. As $r_K(K) < 0$, $r_K(1) > 0$, and the derivative is always positive, there exists exactly one positive root ξ of $r_K(z)$ and it lies in $(K, 1)$. Using the formula for the roots of a cubic polynomial we actually get

$$\xi = -\frac{1}{3}\sqrt[3]{-27K + 3\sqrt{81K^2+3}} - \frac{1}{3}\sqrt[3]{-27K - 3\sqrt{81K^2+3}}.$$

Definition 6.4. Given $K \in (0,1)$, we consider the polynomial $r_K(z) = z^3 + z - 2K$ and we call ξ its unique positive root (Remark 6.3). We let Θ be the following subset of $(0,1)^2$:

$$\Theta = \left\{ (\lambda,\mu) \in (0,1)^2 \middle| \xi + 1 \leq \lambda + \mu < 2,\ |\lambda - \mu| < \min\left\{ 2 - \lambda - \mu, \sqrt{\tfrac{r_K(\lambda+\mu-1)}{\lambda+\mu-1}} \right\} \right\}.$$

Lemma 6.5. *Let λ, μ be real numbers in $(0,1)$ with $\lambda + \mu > 1$. Then (λ, μ) belongs to Θ if and only if*

$$\frac{K}{\lambda + \mu - 1} - (1-\lambda)(1-\mu) < \lambda\mu. \tag{3}$$

Proof. As $\lambda + \mu > 1$, we exchange the inequality (3) by the following equivalent inequality:

$$(\lambda + \mu - 1)(2\lambda\mu + 1 - \lambda - \mu) - K > 0. \tag{4}$$

We consider the change of variables $s := \lambda + \mu$, $t := \lambda - \mu$ (so that $\lambda = \frac{s+t}{2}$, $\mu = \frac{s-t}{2}$). We observe that λ and μ lie in $(0,1)$ if and only if $|t| < s$ and $|t| < 2 - s$. As we are assuming $\lambda + \mu > 1$, we have $s > 2 - s$. Therefore, λ, μ are real numbers in $(0,1)$ with $\lambda + \mu > 1$ if and only if $|t| < 2 - s$.

In these new variables inequality (4) reads as $(s-1)(\frac{s^2-t^2}{2} + 1 - s) - K > 0$, which is equivalent to

$$t^2 < \frac{(s-1)((s-1)^2+1) - 2K}{s-1} = \frac{r_K(s-1)}{s-1}. \tag{5}$$

$\Leftarrow$) Let λ, μ be real numbers in $(0,1)$ satisfying $\lambda + \mu > 1$ and (4). Then $s := \lambda + \mu$ lies in $(1,2)$, $|t := \lambda - \mu| < 2 - s$, and s, t satisfy (5). In particular, $\frac{r_K(s-1)}{s-1} \geq 0$. As we have $s > 1$, this inequality is positive if and only if its

numerator is positive, which holds if and only if $s-1 \geq \xi$. Therefore s is in $[\xi+1, 2)$ and $|t| < \min\left\{2-s, \sqrt{\frac{r_K(s-1)}{s-1}}\right\}$; in other words, (λ,μ) belongs to Θ.

$\Rightarrow$) Conversely, let $(\lambda,\mu) \in \Theta$. Then, using the change of variables above, we have that (s,t) satisfies $|t| < \sqrt{\frac{r_K(s-1)}{s-1}}$. In particular, (5) is satisfied and hence (3) is satisfied as well. □

Proposition 6.6. *Given K a real number in $(0,1)$, we consider the polynomial $r_K(z) = z^3+z-2K$ and let ξ be its positive real root in $(K,1)$ (see Remark 6.3). We fix two real numbers λ,μ in $(0,1)$ such that $\lambda+\mu>1$. Then the set*

$$\Omega_{\lambda,\mu} = \left\{(\alpha,\beta)\in\mathbb{R}^2 \,\middle|\, 0<\alpha<\mu, |\beta|<\lambda, |\alpha\beta - \frac{K}{\lambda+\mu-1}| < (1-\lambda)(1-\mu)\right\}$$

is non-empty if and only if (λ,μ) belongs to Θ. Moreover in this case, (α,β) belongs to $\Omega_{\lambda,\mu}$ if and only if α belongs to $\left(\frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\lambda}, \mu\right)$, $\alpha>0$, and

$$\max\left\{-\lambda, \frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\alpha}\right\} < \beta < \min\left\{\lambda, \frac{\frac{K}{\lambda+\mu-1}+(1-\lambda)(1-\mu)}{\alpha}\right\}.$$

Proof. $\Rightarrow$) If (α,β) is a point in $\Omega_{\lambda,\mu}$, then $|\alpha\beta - \frac{K}{\lambda+\mu-1}| < (1-\lambda)(1-\mu)$. This is equivalent to

$$\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \alpha\beta < \frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu). \tag{6}$$

In particular, as $\alpha\beta < \lambda\mu$, we have

$$\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \lambda\mu.$$

Hence, using Lemma 6.5 we obtain $(\lambda,\mu)\in\Theta$.

Moreover, as $|\beta|<\lambda$, inequality $\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \alpha\beta$ implies $\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu) < \lambda\alpha$, and therefore α belongs to the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\lambda}, \mu\right).$$

The inequalities on β follow directly from (6) and from $|\beta|<\lambda$. Conversely, if α belongs to the above interval, and β satisfies

$$\max\left\{-\lambda, \frac{\frac{K}{\lambda+\mu-1}-(1-\lambda)(1-\mu)}{\alpha}\right\} < \beta < \min\left\{\lambda, \frac{\frac{K}{\lambda+\mu-1}+(1-\lambda)(1-\mu)}{\alpha}\right\},$$

then inequalities (6) hold and hence (α, β) lies in $\Omega_{\lambda,\mu}$.

$\Leftarrow$) Let (λ, μ) be a point in Θ. In this case (λ, μ) satisfies (3), and in particular, the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\lambda}, \mu \right) \tag{7}$$

is non-empty. We choose $\alpha > 0$ in this interval.

Then, the interval

$$\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha}, \frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha} \right)$$

is non-empty (the left-hand side numerator is smaller than the right-hand side numerator, and the denominator is positive) and its intersection with $(-\lambda, \lambda)$ is not empty. Indeed, as $\alpha > 0$ and α belongs to the interval (7), we have

$$\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha} < \lambda;$$

moreover $-\lambda$ is less than $\frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha}$ because this expression is positive.

Finally, we choose β in this intersection of intervals and we obtain a point (α, β) in $\Omega_{\lambda,\mu}$. □

Theorem 6.7. *Let K be a real number in $(0, 1)$.*

(a) *Let (λ, μ) be a point in Θ, let (α, β) be a point in $\Omega_{\lambda,\mu}$, and consider real numbers α' and β' such that*
   (i) $\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu} < |\beta'| < 1 - \lambda$, *and*
   (ii) $\alpha' = \frac{\alpha\beta - \frac{K}{\lambda+\mu-1}}{\beta'}$.
   Then, if we consider the change of variables $a = (\lambda + \beta)/2$,$b = (1 - \lambda - \beta')/2$,$c = (1-\lambda+\beta')/2$ $d = (\lambda-\beta)/2$, $e = (1-\mu-\alpha')/2$, $f = (\mu+\alpha)/2$, $g = (\mu - \alpha)/2$, $h = (1 - \mu + \alpha')/2$, the matrix

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

   is a strictly stochastic SSM *matrix with determinant K, $a+d+f+g > 1$, $b \neq c$, and $f > g$.*

(b) *Conversely, let*

$$A = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ h & g & f & e \\ d & c & b & a \end{pmatrix}$$

be a strictly stochastic SSM *matrix with determinant* K *and with* $a+d+g+f>1$, $b \neq c$ *and* $f > g$. *Then* $F(A)$ *is equal to*

$$\begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix},$$

where $(\lambda, \mu) \in \Theta$, $(\alpha, \beta) \in \Omega_{\lambda,\mu}$, *and* α', β' *satisfy conditions (i) and (ii) stated in* (a).

Remark 6.8. (1) By Proposition 6.6, if (λ, μ) is a point in Θ, there exists $(\alpha, \beta) \in \Omega_{\lambda,\mu}$. This implies that $|\alpha\beta - \frac{K}{\lambda+\mu-1}|$ is smaller than $(1-\lambda)(1-\mu)$, and thus the interval

$$\left(\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu}, 1-\lambda \right)$$

is non-empty. In particular, there exists β' in this interval. Therefore conditions (i) and (ii) in Theorem 6.7(a) are not empty.

(2) Assumptions $a+d+g+f>1$, $f>g$, $b \neq c$ are biologically meaningful: the elements in the diagonal of an evolutionary Markov matrix stand for the conditional probabilities of no mutation, which are supposed to be much higher than the off-diagonal probabilities. It is even reasonable to assume that these diagonal entries are greater than 0.5, giving in particular $a+d+g+f>1$. In any case, the result proved above can be easily adapted to the case $a+d+g+f<1$ or $f>g$ (we have not done it here in order to make the paper more readable). Note also that any SSM matrix with determinant K and $f<g$ gives rise to an SSM matrix with $f>g$ and determinant K by permuting its 1st and 4th rows and its 2nd and 3rd rows (or columns, if preferred).

The hypothesis $b \neq c$ was added to simplify the statement of the Theorem and can be easily removed. Indeed, a matrix A as in (b) has $b=c$ and determinant equal to K if and only if $F(A)$ has $\beta'=0$ and K is equal to $(\lambda+\mu-1)\alpha\beta$. Therefore A is strictly stochastic with determinant K and $b=c$ if and only if $\frac{K}{\lambda(\lambda+\mu-1)} < |\alpha| < \mu$, $\beta = \frac{K}{\alpha(\lambda+\mu-1)}$, $\beta'=0$ and α' is any number satisfying $|\alpha'| < 1-\mu$.

Proof. (a) Let A be defined from λ, μ, $\alpha, \alpha', \beta, \beta'$ as in the statement. Then $F(A)$ is equal to

$$B = \begin{pmatrix} \lambda & 1-\lambda & 0 & 0 \\ 1-\mu & \mu & 0 & 0 \\ 0 & 0 & \alpha & \alpha' \\ 0 & 0 & \beta' & \beta \end{pmatrix}.$$

We prove that A is a stochastic matrix using Lemma 6.2.

By hypothesis, $(\lambda, \mu) \in \Theta$ and hence λ and μ lie in $(0, 1)$. Moreover, as $(\alpha, \beta) \in \Omega_{\lambda,\mu}$, we have $0 < \alpha < \mu$, $|\beta| < \lambda$. By assumption (i), $|\beta'| < 1 - \lambda$ is also satisfied. It remains to prove that $|\alpha'| < 1 - \mu$. But this follows from conditions (i) and (ii):

$$|\alpha'| = \frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{|\beta'|} < 1 - \mu.$$

Row sums in A are equal to 1 by definition of $a, b, \ldots, h$. Moreover, as $B = F(A)$ is obtained from A by a basis change, we have that $\det A = \det B$ and it coincides with $(\lambda + \mu - 1)(\alpha\beta - \alpha'\beta')$. Thus, by assumption (ii) we have $\det A = K$.

(b) Lemma 6.2 tells us that $F(A)$ has the shape in the statement of the theorem, and that $\lambda = a + d$, $\mu = g + f$, $\alpha = f - g$, $\alpha' = h - e$, $\beta = a - d$, and $\beta' = c - b$. Moreover, as A is strictly stochastic, we have that λ, μ lie in $(0, 1)$, $|\alpha| < \mu$, $|\beta| < \lambda$, $|\alpha'| < 1-\mu$, $|\beta'| < 1-\lambda$. We are also assuming $a+d+g+f > 1$, $b \neq c$, and $f > g$, so that we have $\lambda + \mu > 1$, $\beta' \neq 0$, and $0 < \alpha < \mu$.

On the other hand, $\det A = K$ implies $K = (\lambda + \mu - 1)(\alpha\beta - \alpha'\beta')$ and therefore condition (ii) holds.

The remaining inequality in (i),

$$\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu} < |\beta'|,$$

holds because $|\alpha'|$ satisfies (ii) and $|\alpha'| < 1 - \mu$.

We prove now that (α, β) belongs to $\Omega_{\lambda\mu}$, that is,

$$|\alpha\beta - \frac{K}{\lambda + \mu - 1}| < (1 - \lambda)(1 - \mu). \tag{8}$$

We have just seen that $|\beta'|$ satisfies condition (i), so

$$|\alpha\beta - \frac{K}{\lambda + \mu - 1}| < |\beta'|(1 - \mu)$$

and this last term is $< (1 - \lambda)(1 - \mu)$. Therefore (8) is satisfied.

Finally, as (α, β) is a point in $\Omega_{\lambda,\mu}$, this set is not empty and (λ, μ) belongs to Θ by Proposition 6.6. □

The previous results and their proofs provide the following algorithm for generating any SSM matrix

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

with $a + d + g + f > 1$, $f > g$, and $b \neq c$.

Algorithm 6.9. (Generation of SSM matrices with given determinant.)

Input: K in $(0, 1)$.

Output: A strictly stochastic SSM matrix A with determinant K.

Step 1: Compute the unique positive root ξ of $r_K(z)$ following Remark 6.3.
Step 2: Take s randomly in $[\xi + 1, 2)$.
Step 3: Take t randomly such that $|t| < \min\left\{2 - s, \sqrt{\frac{r_K(s-1)}{s-1}}\right\}$.
Step 4: Set $\lambda = \frac{s+t}{2}$ and $\mu = \frac{s-t}{2}$.
Step 5: Take $\alpha > 0$ randomly in $\left(\frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\lambda}, \mu\right)$.
Step 6: Choose β randomly such that

$$\max\left\{-\lambda, \frac{\frac{K}{\lambda+\mu-1} - (1-\lambda)(1-\mu)}{\alpha}\right\} < \beta < \min\left\{\lambda, \frac{\frac{K}{\lambda+\mu-1} + (1-\lambda)(1-\mu)}{\alpha}\right\}.$$

Step 7: Choose β' randomly such that $\frac{|\alpha\beta - \frac{K}{\lambda+\mu-1}|}{1-\mu} < |\beta'| < 1 - \lambda$.
Step 8: Set $\alpha' := \frac{\alpha\beta - \frac{K}{\lambda+\mu-1}}{\beta'}$, $a := (\lambda + \beta)/2$,$b := (1 - \lambda - \beta')/2$,$c := (1 - \lambda + \beta')/2$ $d := (\lambda - \beta)/2$, $e := (1 - \mu - \alpha')/2$, $f := (\mu + \alpha)/2$, $g := (\mu - \alpha)/2$, and $h := (1 - \mu + \alpha')/2$.
Final: Return

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

Remark 6.10. As SSM matrices include K81* matrices, using Remark 5.5 we see that there exist matrices produced by the algorithm above that are not of type $\exp(Q)$ for a real matrix Q.

7. Generating GMM matrices with given determinant

For GMM matrices we do not have such a general result as in the previous sections. We do not know how to generate *any* strictly stochastic GMM matrix, but here we explain a way for generating some of them.

We could obtain a strictly stochastic matrix GMM matrix with determinant equal to K by exponentiating a rate matrix (i.e. a matrix with row sums equal to 0 and off-diagonal positive entries) with trace equal to $\log K$ (cf. [15, Theorem 4.19]). However, not all GMM matrices are of this type (see [7] and Remark 5.5). We use the fact that the product of two strictly stochastic matrices is again a strictly stochastic matrix in order to obtain a broader class of GMM matrices. In fact, we multiply a GMM matrix of type $\exp(Q)$ with determinant $\delta > K$ by an SSM matrix of determinant K/δ. We must admit that we do not know how much larger is this class of matrices. The set V of GMM matrices with determinant K corresponds to an affine variety of dimension 11. There are 11 free parameters for a rate matrix Q with given trace, so the matrices of type $\exp(Q)$ lie on a subset of V of dimension 11. Therefore the set of matrices produced by the algorithm below form a subset of maximum dimension of V, and this subset is larger than the set $\{\exp(Q) | Q \text{ rate matrix, } \operatorname{tr} Q = K\}$.

Algorithm 7.1. (Generation of GMM matrices with given determinant.)

Input: K in $(0, 1)$.

Output: A strictly stochastic GMM matrix A with determinant K.

Step 1: Take a random number t in $(\log K, 0)$.
Step 2: Generate a random rate matrix Q with nonzero entries and $\operatorname{tr} Q = t$.
Step 3: Compute $A_0 = \exp(Q)$.
Step 4: Following algorithm 6.9, generate a strictly stochastic SSM matrix B with determinant equal to K/e^t.
Final: Return $A = BA_0$.

Remark 7.2. In the above algorithm, even if we considered matrices B with real logarithm ($B = \exp(R)$), the matrices $A = BA_0$ produced above do not need to be of exponential type.

Acknowledgments

We thank the referee for useful comments. Both authors are partially supported by Generalitat de Catalunya, 2009 SGR 1284. Research of the first author is partially supported by Ministerio de Educación y Ciencia MTM2009-14163-C02-02 and MTM2012-38122-C03-01.

References

[1] ES Allman and JA Rhodes, *Phylogenetic invariants for the general Markov model of sequence mutation*, Mathematical Biosciences **186** (2003), no. 2, 113–144. MR MR2024609 (2004j:92048)

[2] ———, *Mathematical models in biology, an introduction*, Cambridge University Press, January 2004, ISBN 0-521-52586-1).

[3] D Barry and JA Hartigan, *Statistical analysis of hominoid molecular evolution*, Statistical Sciences **2** (1987), no. 2, 191–207.

[4] M Casanellas and J Fernandez-Sanchez, *Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees*, Mol. Biol. Evol. **24** (2007), no. 1, 288–293.

[5] ———, *Relevant phylogenetic invariants of evolutionary models*, Journal de Mathématiques Pures et Appliquées **96** (2011), 207–229.

[6] M Casanellas and S Sullivant, *The strand symmetric model*, Algebraic Statistics for computational biology (L. Pachter and B. Sturmfels, eds.), Cambridge University Press, 2005.

[7] W J Culver, *On the existence and uniqueness of the real logarithm of a matrix*, Proc. Amer. Math. Soc. **17** (1966), 1146–1151.

[8] J Draisma and J Kuttler, *On the ideals of equivariants tree models*, Mathematische Annalen **344** (2009), 619–644.

[9] M Hasegawa, H Kishino, and T Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*, Journal of Molecular Evolution **22** (1985), 160–174.

[10] T H Jukes and C R Cantor, *Evolution of protein molecules*, Mammalian Protein Metabolism (H N Munro, ed.), vol. 3, Academic Press, 1969, pp. 21–132.

[11] A Kedzierska and M Casanellas, *`GenNon-h`: Generating multiple sequence alignments on nonhomogeneous phylogenetic trees*, BMC Bioinformatics **13** (2012), 216.

[12] A Kedzierska, M Drton, R Guigó, and M Casanellas, *`SPIn`: model selection for phylogenetic mixtures via linear invariants*, Molecular Biology and Evolution **29** (2012), 929–937.

[13] M Kimura, *A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences*, Journal of Molecular Evolution **16** (1980), 111–120.

[14] ———, *Estimation of evolutionary sequences between homologous nucleotide sequences*, Proceedings of the National Academy of Sciences **78** (1981), 454–458.

[15] L Pachter and B Sturmfels (eds.), *Algebraic statistics for computational biology*, Cambridge University Press, November 2005, ISBN 0-521-85700-7.

[16] A Rambaut and NC Grassly, *Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees*, Comput. Appl. Biosci. **13** (1997), 235–238.

[17] C Semple and M Steel, *Phylogenetics*, Oxford Lecture Series in Mathematics and its Applications, vol. 24, Oxford University Press, Oxford, 2003.

[18] MA Steel, *Recovering a tree from the leaf colourations it generates under a markov model*, Applied Mathematics Letters **7** (1994), 19–24.

[19] B Sturmfels and S Sullivant, *Toric ideals of phylogenetic invariants*, Journal of Computational Biology **12** (2005), 204–228.

[20] S Tavaré, *Some probabilistic and statistical problems in the analysis of DNA sequences*, Some mathematical questions in biology—DNA sequence analysis (New York, 1984), Lectures Math. Life Sci., vol. 17, Amer. Math. Soc., Providence, RI, 1986, pp. 57–86.

[21] Z Yang, *PAML: A program package for phylogenetic analysis by maximum likelihood*, CABIOS **15** (1997), 555–556.

Departament de Matemàtica Aplicada I. ETSEIB. Universitat Politècnica de Catalunya. Avinguda Diagonal 647. 08028 Barcelona. Spain.

E-mail address: marta.casanellas@upc.edu

Centre for Genomic Regulation (CRG). Dr. Aiguader 88. 08003 Barcelona. Spain.

E-mail address: anna.kedzierska@upc.edu