# Supplementary Material for "Distinguishing Between Convergent Evolution and Violation of the Molecular Clock for Three Taxa" 

Jonathan D. Mitchell ${ }^{1,2}$, Jeremy G. Sumner ${ }^{1}$, and Barbara R. Holland ${ }^{1}$<br>${ }^{1}$ School of Natural Sciences, University of Tasmania, Hobart, Tasmania 7001, Australia;<br>${ }^{2}$ Department of Mathematics $\mathcal{B}$ Statistics, University of Alaska Fairbanks, Alaska 99775, USA

Corresponding author: Jonathan D. Mitchell, Department of Mathematics \& Statistics, University of Alaska Fairbanks, Alaska 99775, USA; E-mail: jdmitchell5@alaska.edu.

## Appendix

The splitting operator splits a single ancestral lineage into two descendant lineages. Initially after splitting no divergence of lineages has occurred and the character patterns with mismatches must have zero probability. Independent evolution after a splitting event, leading to divergence of lineages, results in the character patterns with mismatches having non-zero probability over time.

To illustrate how the splitting operator works mathematically, first consider the two unit vectors for a two-state Markov model, such as the binary symmetric model. The unit vectors corresponding to the two character states for a single lineage are

$$
\left.e_{0}=\begin{array}{cc}
0 & 1 \\
{\left[\begin{array}{ll}
1 & 0
\end{array}\right]}
\end{array} \quad \begin{array}{cc}
0 & 1 \\
\text { and } & e_{1}= \\
0 & 1
\end{array}\right]
$$

where the labels on the columns of the vectors refer to the two possible states: 0 or 1 , for a single lineage. The unit vectors are expressed as row vectors to be consistent with row sum convention for transition matrices and rate matrices. Throughout the appendix rows of transition matrices will sum to 1 and rows of rate matrices will sum to 0 .

Character pattern probabilities for a single lineage can be expressed as $\left[\begin{array}{ll}p_{0} & p_{1}\end{array}\right]=p_{0} e_{0}+p_{1} e_{1}$. The splitting operator acts on the two unit vectors to produce two new unit vectors representing two lineages. The splitting operator performs the action of
the matrix Kronecker product of a unit vector for a single lineage with itself:

$$
\left\{\begin{array}{c}
\delta\left(e_{0}\right)=e_{0} \otimes e_{0}=\left[\begin{array}{llll}
00 & 01 & 10 & 11 \\
1 & 0 & 0 & 0
\end{array}\right]=e_{00} \\
\delta\left(e_{1}\right)=e_{1} \otimes e_{1}=\left[\begin{array}{llll}
00 & 01 & 10 & 11 \\
0 & 0 & 0 & 1
\end{array}\right]=e_{11}
\end{array}\right.
$$

where $\otimes$ is the matrix Kronecker product, and columns are labeled by character patterns.
Acting on the two unit vectors for a single lineage, the splitting operator produces only two of the four unit vectors for two lineage: $e_{00}$ and $e_{11}$. These are the two unit vectors corresponding to the combinations of characters where the two lineages are the same: 00 and 11. Initially after the splitting event the only possible combinations of characters are 00 and 11. If the two lineages were to then diverge all possible combinations of characters: $00,01,10$ and 11 , could occur. Character pattern probabilities would then be expressed in terms of all four unit vectors: $e_{00}, e_{01}, e_{10}$ and $e_{11}$.

Consider the 2-taxon clock-like tree in Figure [1. Suppose

$$
\Pi=\left[\begin{array}{cc}
0 & 1 \\
\frac{1}{2} & \frac{1}{2}
\end{array}\right]
$$

is the root distribution for the single lineage present at the root. This is the stationary distribution, the long run probability distribution for a single lineage. Immediately after the root the single lineage evolves according to the $2 \times 2$ transition matrix $M$. After some period of time the splitting operator $\delta$ then splits the single lineage into two descendant lineages, which then diverge. The vector of character pattern probabilities immediately after the split is $P=\Pi \cdot M \cdot \delta$, where $\delta$ is now being expressed as a matrix and $\cdot$ is the
matrix multiplication.


Figure 1: A 2-taxon clock-like tree. (Top) In Figure ${ }^{\text {Da }}$ the stationary distribution $\Pi$ is at the root, a Markov matrix $M$ acts on the single descendant lineage and the splitting operator $\delta$ splits the lineage into two lineages after some period of time after the root. Figure is an equivalent 2 -taxon clock-like tree where the two lineages split immediately after the root and remain identical for a period of time. The period where the two lineages have split but remain identical is represented by the transition matrix $\mathcal{M}$. (Bottom) State transition diagrams for the epoch between the dashed lines of each subfigure. Dark gray nodes indicate possible states during the epoch, while light gray nodes indicate impossible states. Solid lines indicate "regular" transitions, dotted lines in Figure indicate transitions that cannot occur since they involve impossible states.

For the binary symmetric model the rate matrix is

$$
\left.Q=\begin{array}{c}
0 \\
1
\end{array} \begin{array}{cc}
0 & 1 \\
-\alpha & \alpha \\
\alpha & -\alpha
\end{array}\right],
$$

and the corresponding transition matrix along an edge of length $t$ is

$$
M=e^{Q t}=\begin{array}{cc}
0 & 1 \\
1
\end{array}\left[\begin{array}{cc}
0 \\
\frac{1}{2}\left(1+e^{-2 \alpha t}\right) & \frac{1}{2}\left(1-e^{-2 \alpha t}\right) \\
\frac{1}{2}\left(1-e^{-2 \alpha t}\right) & \frac{1}{2}\left(1+e^{-2 \alpha t}\right)
\end{array}\right] .
$$

In matrix form the splitting operator for a two-state Markov model is

$$
\left.\delta=\begin{array}{c}
0 \\
1
\end{array} \begin{array}{llll}
00 & 01 & 10 & 11 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1
\end{array}\right] .
$$

The character pattern probabilities immediately after the split are then

$$
P=\Pi \cdot M \cdot \delta=\left[\begin{array}{cccc}
00 & 01 & 10 & 11 \\
\frac{1}{2} & 0 & 0 & \frac{1}{2}
\end{array}\right],
$$

since $\Pi \cdot M=\Pi$ for the stationary distribution.
An equivalent representation of the 2-taxon clock-like tree in question is that of Figure [1]. Rather than splitting after some period of time, the splitting operator can be "pushed back" to the root and two lineages can be modeled as having split at the root and then remained identical (identical character pattern probabilities) for a period of time.

Note that for $n$ taxa, with $n-1$ splitting operators, all splitting operators can be "pushed back" to the root. Pushing the $n-1$ splitting operators back to the root then gives a root distribution for $n$ lineages. The two equivalent representations of the character pattern probabilities at the time of the split for the 2-taxon clock-like tree are $P=\Pi \cdot M \cdot \delta=\Pi \cdot \delta \cdot \mathcal{M}$, where $\mathcal{M}$ is a $4 \times 4$ transition matrix acting on two identical lineages to keep them identical. The rate matrix $\mathcal{Q}$ corresponding to this transition matrix
such that $\mathcal{M}=e^{\mathcal{Q} t}$ is

$$
\mathcal{Q}=\begin{gather*}
01  \tag{1}\\
01 \\
10 \\
11
\end{gather*}\left[\begin{array}{cccc}
00 & 01 & 10 & 11 \\
-\alpha & 0 & 0 & \alpha \\
\alpha & -2 \alpha & 0 & \alpha \\
\alpha & 0 & -2 \alpha & \alpha \\
\alpha & 0 & 0 & -\alpha
\end{array}\right] .
$$

Below we show how to form a general rate matrix $\mathcal{Q}$.
Note that if the only possible initial character patterns are 00 and 11 then the only permissible substitutions are between these two character patterns, and 01 and 10 cannot be achieved. Two lineages that are originally identical will remain identical.

However, if divergence of the two lineages has occurred then after the divergence the character patterns 01 and 10 will have non-zero probability of occurring. With the rate matrix $\mathcal{Q}$ the only permissible substitutions from character patterns 01 and 10 are to character patterns 00 and 11 . If the rate matrix $\mathcal{Q}$ is then applied to the two diverged lineages it will act to model "convergence", where the two lineages become more similar over time; in the limit they would become identical.

In general, a single rate matrix can be formed that models $k$ edges during some epoch, some of which are involved in convergence and some of which are not, with potentially multiple convergence processes occurring at the same time. For example, a single rate matrix could be used to model the processes in each epoch of Figure [2], including the third epoch of Figure 2 Zc .

A partition $\mathbb{P}$ of the lineages can be formed that represents all convergence blocks of lineages. Each element of the partition consists of a block of lineages that are all converging together. Lineages that are diverging from all others are in their own "convergence" block. For example, the partition representing taxa 1 and 2 converging and taxa 3 diverging in


Figure 2: Three epochs of a 3-taxon process. (Top) Divergence of lineages is represented by straight lines emanating from a node, while convergence in the third epoch in Figure [20 is represented by curved lines. (Bottom) State transition diagrams for the last epoch of each subfigure. Dark gray nodes indicate possible states during the epoch, while light gray nodes indicate impossible states. In Figure 2 impossible states are those where lineages 2 and 3 are not identical. Solid lines indicate "regular" transitions, dotted lines in Figure Za indicate transitions that cannot occur since they involve impossible states and dashed lines indicate "correction" transitions responsible for convergence in Figure 2 Cl .

Figure $\mathbb{Z D}$ is $\mathbb{P}=\{\{1,2\},\{3\}\}$. Note that lineages can be involved in multiple convergence blocks and thus may not necessarily be diverging from all other lineages not in a particular convergence block. For example, a potential partition for a 3-taxon convergence-divergence model could be $\mathbb{P}=\{\{1,2\},\{2,3\}\}$, where taxa 1 and 2 are converging together and taxa 2
and 3 are converging together. This is different to the partition $\mathbb{P}=\{\{1,2,3\}\}$, where all three taxa are converging together, which would be represented by a different rate matrix.

In general, given a partition $\mathbb{P}$ of the lineages associated with a particular epoch we construct the rate matrix $\mathcal{Q}_{\mathbb{P}}$ that is associated with the epoch as follows. For each block (convergence group or individual diverging lineage) $B \in \mathbb{P}$ of the partition we construct a component $\mathbb{R}_{B}^{(n)}$ of the rate matrix $\mathcal{Q}_{\mathbb{P}}$ :

$$
\mathbb{R}_{B}^{(n)}=\sum_{x \in P(B) \backslash \emptyset} R_{x, \alpha}^{(n)}+R_{x, \beta}^{(n)},
$$

where $n$ is the number of taxa, $P(B)$ is the power set of $B, \emptyset$ is the empty set, $R_{x, y}^{(n)}=S_{1, y} \otimes S_{2, y} \otimes \ldots \otimes S_{n, y}$, where $\otimes$ is the matrix Kronecker product and

$$
S_{i, y}=\left\{\begin{array}{cl}
L_{\alpha}=\left[\begin{array}{cc}
-1 & 1 \\
0 & 0
\end{array}\right] \quad \text { if } y=\alpha \text { and } i \in x \\
L_{\beta}=\left[\begin{array}{cc}
0 & 0 \\
1 & -1
\end{array}\right] & \text { if } y=\beta \text { and } i \in x \\
I=\left[\begin{array}{cc}
1 & 0 \\
0 & 1
\end{array}\right] \quad \text { otherwise. }
\end{array}\right.
$$

Finally, $\mathcal{Q}_{\mathbb{P}}^{(n)}=\alpha \sum_{B \in \mathbb{P}} \mathbb{R}_{B}^{(n)}$.
For example, for $n=2$, the rate matrix of Equation $\mathbb{D}$ is
$\mathcal{Q}_{\{\{1,2\}\}}^{(2)}=\alpha \mathbb{R}_{\{1,2\}}^{(2)}=\alpha\left(I \otimes L_{\alpha}+L_{\alpha} \otimes I+L_{\alpha} \otimes L_{\alpha}+I \otimes L_{\beta}+L_{\beta} \otimes I+L_{\beta} \otimes L_{\beta}\right)$. For another example, suppose $n=3$. Then for the third epoch of Figure [2d, one of the components of the rate matrix is
$\mathbb{R}_{\{1,2\}}^{(3)}=I \otimes L_{\alpha} \otimes I+L_{\alpha} \otimes I \otimes I+L_{\alpha} \otimes L_{\alpha} \otimes I+I \otimes L_{\beta} \otimes I+L_{\beta} \otimes I \otimes I+L_{\beta} \otimes L_{\beta} \otimes I$. The rate matrix is the sum of the components relating to each block of the partition
multiplied by the substitution rate. The rate matrix for the third epoch of Figure [2d with the first two taxa converging and the third taxon diverging is

$$
\left.\begin{array}{r}
000 \\
\mathcal{Q}_{\{\{1,2\},\{3\}\}}^{(3)}=\alpha\left(\mathbb{R}_{\{1,2\}}^{(3)}+\mathbb{R}_{\{3\}}^{(3)}\right)= \\
001 \\
0100
\end{array} \begin{array}{cccccccc}
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
-2 \alpha & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 \\
\alpha & -2 \alpha & 0 & 0 & 0 & 0 & 0 & \alpha \\
\alpha & 0 & -3 \alpha & \alpha & 0 & 0 & \alpha & 0 \\
0 & \alpha & \alpha & -3 \alpha & 0 & 0 & 0 & \alpha \\
\alpha & 0 & 0 & 0 & -3 \alpha & \alpha & \alpha & 0 \\
0 & \alpha & 0 & 0 & \alpha & -3 \alpha & 0 & \alpha \\
\alpha & 0 & 0 & 0 & 0 & 0 & -2 \alpha & \alpha \\
0 & \alpha & 0 & 0 & 0 & 0 & \alpha & -2 \alpha
\end{array}\right] .
$$

Note that $\mathcal{Q}_{\{\{1,2\},\{3\}\}}^{(3)}$ does not permit any substitutions where taxa 1 and 2 become less similar. That is, any substitutions where taxa 1 and 2 start in the same state and end in different states are not permitted.

The vector for the character pattern probabilities is then
$P=\Pi \cdot \delta^{(2)} \cdot \mathcal{M}_{\{\{1\},\{2,3\}\}} \cdot \mathcal{M}_{\{\{1\},\{2\},\{3\}\}} \cdot \mathcal{M}_{\{\{1,2\},\{3\}\}}$, where $\mathcal{M}_{\mathbb{P}}=e^{\mathcal{Q}_{\mathbb{P}} t_{\mathbb{P}}}, t_{\mathbb{P}}$ is the epoch time and the superscript for the number of taxa has been dropped for simplicity. Note that

$$
\left.\delta^{(2)}=\begin{array}{c}
0 \\
1
\end{array} \begin{array}{cccccccc}
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]
$$

performs the action of splitting the root distribution for one lineage to three lineages as
follows:

$$
\Pi \cdot \delta^{(2)}=\left[\begin{array}{cccccccc}
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2}
\end{array}\right]
$$

