

# Identifying Convergence in Phylogenetic Models

Jonathan Mitchell

Supervisors: Barbara Holland, Jeremy Sumner

University of Tasmania

November 11, 2011

# Introduction

- In this presentation I will look at the probability distributions for various two and three taxa trees.

# Introduction

- In this presentation I will look at the probability distributions for various two and three taxa trees.
- For each number of taxa, I will compare trees with convergence to trees without convergence.

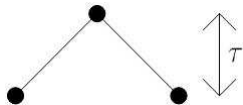
# Introduction

- In this presentation I will look at the probability distributions for various two and three taxa trees.
- For each number of taxa, I will compare trees with convergence to trees without convergence.
- I will look at whether the case of convergence can be distinguished from the case of no convergence.

# Introduction

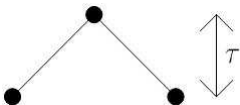
- In this presentation I will look at the probability distributions for various two and three taxa trees.
- For each number of taxa, I will compare trees with convergence to trees without convergence.
- I will look at whether the case of convergence can be distinguished from the case of no convergence.
- I will apply the binary symmetric model to each tree for simplicity.

## Clock-Like Two Taxa Tree



- Binary symmetric model for evolution.

## Clock-Like Two Taxa Tree



- Binary symmetric model for evolution.
- Probability distribution for a clock-like tree is

$$P(\tau) = \begin{bmatrix} p_{00}(\tau) \\ p_{01}(\tau) \\ p_{10}(\tau) \\ p_{11}(\tau) \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 + e^{-\tau} \\ 1 - e^{-\tau} \\ 1 - e^{-\tau} \\ 1 + e^{-\tau} \end{bmatrix} .$$

## Transforming the Basis

- Want to express the probability distribution in a more appropriate basis.



## Transforming the Basis

- Want to express the probability distribution in a more appropriate basis.
- Hadamard basis will express the probability distribution as a set of simple constraints.

## Transforming the Basis

- Want to express the probability distribution in a more appropriate basis.
- Hadamard basis will express the probability distribution as a set of simple constraints.
- Transformed probability distribution is

$$\begin{aligned}\hat{P}(\tau) &= \begin{bmatrix} q_{00}(\tau) \\ q_{01}(\tau) \\ q_{10}(\tau) \\ q_{11}(\tau) \end{bmatrix} = \begin{bmatrix} p_{00}(\tau) + p_{01}(\tau) + p_{10}(\tau) + p_{11}(\tau) \\ p_{00}(\tau) - p_{01}(\tau) + p_{10}(\tau) - p_{11}(\tau) \\ p_{00}(\tau) + p_{01}(\tau) - p_{10}(\tau) - p_{11}(\tau) \\ p_{00}(\tau) - p_{01}(\tau) - p_{10}(\tau) + p_{11}(\tau) \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-\tau} \end{bmatrix}.\end{aligned}$$

## Transforming the Basis

- Desirable basis because all of the transformed probabilities have simple and meaningful expressions.

## Transforming the Basis

- Desirable basis because all of the transformed probabilities have simple and meaningful expressions.
- For example, the first constraint tells us that probability must be conserved.

## Transforming the Basis

- Desirable basis because all of the transformed probabilities have simple and meaningful expressions.
- For example, the first constraint tells us that probability must be conserved.
- Last constraint must lie within the range

$$0 < e^{-\tau} \leq 1.$$

## Transforming the Basis

- Desirable basis because all of the transformed probabilities have simple and meaningful expressions.
- For example, the first constraint tells us that probability must be conserved.
- Last constraint must lie within the range

$$0 < e^{-\tau} \leq 1.$$

- Constraints for the probability distribution to have arisen on the tree are

$$\left\{ \begin{array}{l} q_{00} = 1 \\ q_{01} = 0 \\ q_{10} = 0 \\ 0 < q_{11} \leq 1 \end{array} \right.$$

# The Network Model

- We need a way to model convergence.

## The Network Model

- We need a way to model convergence.
- Sumner, Holland and Jarvis proposed the network model for converging taxa.



## The Network Model

- We need a way to model convergence.
- Sumner, Holland and Jarvis proposed the network model for converging taxa.
- Rate matrix for network model is

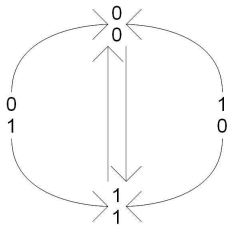
$$Q = \left[ \begin{array}{c|cccc} & 00 & 01 & 10 & 11 \\ \hline 00 & -\alpha & \alpha & \alpha & \alpha \\ 01 & 0 & -2\alpha & 0 & 0 \\ 10 & 0 & 0 & -2\alpha & 0 \\ 11 & \alpha & \alpha & \alpha & -\alpha \end{array} \right].$$

## The Network Model

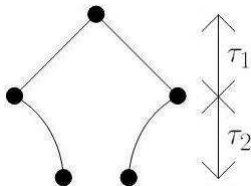
- We need a way to model convergence.
- Sumner, Holland and Jarvis proposed the network model for converging taxa.
- Rate matrix for network model is

$$Q = \left[ \begin{array}{c|cccc} & 00 & 01 & 10 & 11 \\ \hline 00 & -\alpha & \alpha & \alpha & \alpha \\ 01 & 0 & -2\alpha & 0 & 0 \\ 10 & 0 & 0 & -2\alpha & 0 \\ 11 & \alpha & \alpha & \alpha & -\alpha \end{array} \right].$$

- This can be expressed diagrammatically as



## Convergence Two Taxa Tree



- Using the network model, the probability distribution is

$$\hat{P}(\tau) = \begin{bmatrix} q_{00}(\tau) \\ q_{01}(\tau) \\ q_{10}(\tau) \\ q_{11}(\tau) \end{bmatrix} = \begin{bmatrix} p_{00}(\tau) + p_{01}(\tau) + p_{10}(\tau) + p_{11}(\tau) \\ p_{00}(\tau) - p_{01}(\tau) + p_{10}(\tau) - p_{11}(\tau) \\ p_{00}(\tau) + p_{01}(\tau) - p_{10}(\tau) - p_{11}(\tau) \\ p_{00}(\tau) - p_{01}(\tau) - p_{10}(\tau) + p_{11}(\tau) \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 - e^{-\tau_2}(1 - e^{-2\tau_1}) \end{bmatrix}.$$

## Convergence Two Taxa Tree

- Last constraint must lie within the range

$$0 < 1 - e^{-\tau_2}(1 - e^{-2\tau_1}) \leq 1.$$

## Convergence Two Taxa Tree

- Last constraint must lie within the range

$$0 < 1 - e^{-\tau_2}(1 - e^{-2\tau_1}) \leq 1.$$

- Constraints are equivalent for both trees.

## Convergence Two Taxa Tree

- Last constraint must lie within the range

$$0 < 1 - e^{-\tau_2}(1 - e^{-2\tau_1}) \leq 1.$$

- Constraints are equivalent for both trees.
- Under the binary symmetric and network models, the two trees cannot be distinguished from each other.

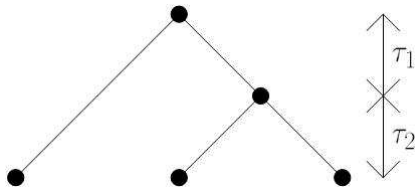
## Convergence Two Taxa Tree

- Last constraint must lie within the range

$$0 < 1 - e^{-\tau_2}(1 - e^{-2\tau_1}) \leq 1.$$

- Constraints are equivalent for both trees.
- Under the binary symmetric and network models, the two trees cannot be distinguished from each other.
- The situation where two diverged taxa begin to converge cannot be distinguished from the case where only divergence has occurred.

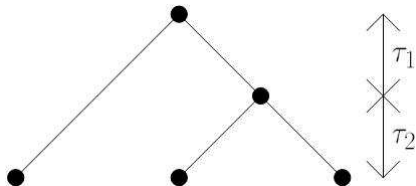
## Clock-Like Three Taxa Tree



- Binary symmetric model for evolution.



## Clock-Like Three Taxa Tree



- Binary symmetric model for evolution.
- Probability distribution for a clock-like tree is

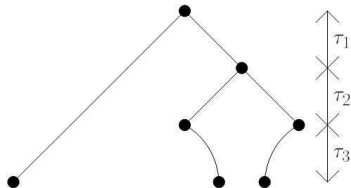
$$\hat{P}(\tau) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-\tau_2} \\ 0 \\ e^{-(\tau_1+\tau_2)} \\ e^{-(\tau_1+\tau_2)} \\ 0 \end{bmatrix} .$$

## Clock-Like Three Taxa Tree

- Constraints for the probability distribution to have arisen on the tree are

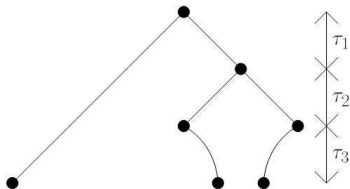
$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

## Convergence Three Taxa Tree



- Binary symmetric model for evolution.

## Convergence Three Taxa Tree



- Binary symmetric model for evolution.
- Probability distribution for a clock-like tree is

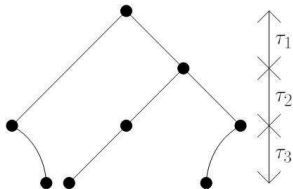
$$\hat{P}(\tau) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-\frac{1}{2}(2\tau_2 + \tau_3)}(1 + e^{\tau_2}(e^{\frac{1}{2}\tau_3} - 1)) \\ 0 \\ e^{-(\tau_1 + \tau_2 + \tau_3)} \\ e^{-(\tau_1 + \tau_2 + \tau_3)} \\ 0 \end{bmatrix} .$$

## Convergence Three Taxa Tree

- Constraints for the probability distribution to have arisen on the tree are

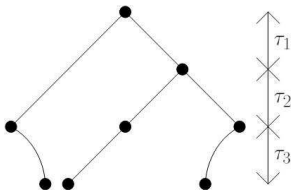
$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

## Convergence Three Taxa Tree



- Binary symmetric model for evolution.

## Convergence Three Taxa Tree



- Binary symmetric model for evolution.
- Probability distribution for a clock-like tree is

$$\hat{P}(\tau) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-(\tau_2+\tau_3)} \\ 0 \\ e^{-(\tau_1+\tau_2+\tau_3)} \\ 1 - e^{-\frac{1}{2}\tau_3}(1 - e^{-(\tau_1+\tau_2)}) \\ 0 \end{bmatrix} \cdot$$

## Convergence Three Taxa Tree

- Constraints for the probability distribution to have arisen on the tree are

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$



## Convergence Three Taxa Tree

- For all three cases:

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

## Convergence Three Taxa Tree

- For all three cases:

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

- For the first two cases:

$$q_{101} = q_{110} \neq q_{011}.$$

## Convergence Three Taxa Tree

- For all three cases:

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

- For the first two cases:

$$q_{101} = q_{110} \neq q_{011}.$$

- For the third case:

$$\left\{ \begin{array}{l} q_{011} \neq q_{101} \\ q_{011} \neq q_{110} \\ q_{101} \neq q_{110} \end{array} \right.$$

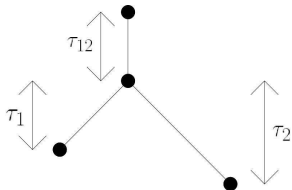
## Convergence Three Taxa Tree

- If the two non-sister taxa begin to converge, this is identifiable from the no convergence case.

## Convergence Three Taxa Tree

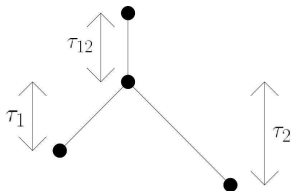
- If the two non-sister taxa begin to converge, this is identifiable from the no convergence case.
- If the two sister taxa begin to converge this is not identifiable from the no convergence case.

## Non Clock-Like Three Taxa Tree



- Binary symmetric model for evolution.

## Non Clock-Like Three Taxa Tree



- Binary symmetric model for evolution.
- Probability distribution for a non clock-like tree is

$$\hat{P}(\tau) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ e^{-(\tau_1 + \tau_2)} \\ 0 \\ e^{-(\tau_2 + \tau_{12})} \\ e^{-(\tau_1 + \tau_{12})} \\ 0 \end{bmatrix}.$$

## Non Clock-Like Three Taxa Tree

- Constraints for the probability distribution to have arisen on the tree are

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$



## Non Clock-Like Three Taxa Tree

- Constraints for the probability distribution to have arisen on the tree are

$$\left\{ \begin{array}{l} q_{000}(\tau) = 1 \\ q_{001}(\tau) = 0 \\ q_{010}(\tau) = 0 \\ 0 < q_{011}(\tau) \leq 1 \\ q_{100}(\tau) = 0 \\ 0 < q_{101}(\tau) \leq 1 \\ 0 < q_{110}(\tau) \leq 1 \\ q_{111}(\tau) = 0 \end{array} \right.$$

- Also have

$$\left\{ \begin{array}{l} q_{011} \neq q_{101} \\ q_{011} \neq q_{110} \\ q_{101} \neq q_{110} \end{array} \right.$$

## Non Clock-Like Three Taxa Tree

- A bijection can be found between the parameters in the tree with non-sister taxa converging and the parameters in the non clock-like tree.

## Non Clock-Like Three Taxa Tree

- A bijection can be found between the parameters in the tree with non-sister taxa converging and the parameters in the non clock-like tree.
- The non clock-like tree cannot be distinguished from the clock-like tree where the two non-sister taxa converge.

## Non Clock-Like Three Taxa Tree

- A bijection can be found between the parameters in the tree with non-sister taxa converging and the parameters in the non clock-like tree.
- The non clock-like tree cannot be distinguished from the clock-like tree where the two non-sister taxa converge.
- The non clock-like tree can be distinguished from all the other trees studied.