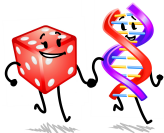


Matrix-analytic methods for the gene-tree species-tree reconciliation problem

Małgorzata O'Reilly ^{a,c} Barbara Holland ^{b,c}

with: Stochastic Modelling Meets Phylogenetics Group



^aUniversity of Tasmania, Australia, malgorzata.oreilly@utas.edu.au

^bUniversity of Tasmania, Australia, barbara.holland@utas.edu.au



^c *This research is funded through Australian Research Council Discovery Project DP180100352.*

25-27 November 2020, Phylomania #12, The Twelfth Theoretical Phylogenetics Meeting at UTAS

Outline

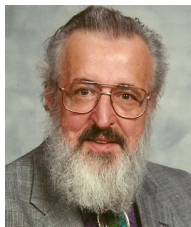
- 1 Introduction
- 2 Species tree¹
- 3 Gene tree¹
- 4 Conclusion

Motivation

- **Matrix-analytic methods (MAMs)** \implies Applications in evolution.
- We describe a model for the **species** , and derive theoretical and algorithmic results for its key measures.
- We describe a model for the **gene** , and derive theoretical and algorithmic results for its key measures.
- We derive MAMs results for the probabilistic analysis of **reconciliation** and illustrate the theory with numerical examples.
- **All code will be made publicly available.**
- *Cite: Results presented here have been derived in ¹.*

¹M.M. O'Reilly et al. Matrix-analytic methods for the gene-tree species-tree reconciliation problem. *To be submitted.*

Matrix-analytic methods (MAMs)



Professor Marcel Neuts (1935-2014)

“His transformative idea was that, rather than developing mathematical structures that have little use for practical applications, the focus should be on constructing models and methods of analysis that can be applied efficiently, using fast algorithms and computers.”²

²B.R. Holland and M.M. O’Reilly. Matrix-analytic methods: Stochastic models for the real world, AMSI Research Report, 2018-2019.

Stochastic Modelling Meets Phylogenetics Group



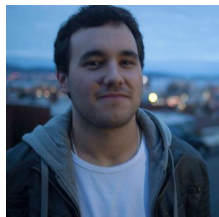
Malgorzata O'Reilly



Barbara Holland



David Liberles



Tristan Stark



Jiahao Diao



Albert Soewengsono



Amanda Wilson

● Mechanistic Models ³

- Mechanistic models are required that are rooted in real life evolutionary/genetic processes.
- Models work well when supported by mechanistic interpretations.
- Need for developing a theoretical science instead of a purely data-driven one.
- Mechanistic models, which provide an explanation of observations, can offer many predictions.

³D.A. Liberles, A.I. Teufel, L. Liu and T. Stadler. On the need for mechanistic models in computational genomics and metagenomics. *Genome Biology and Evolution*, 5(10):2008–2018, 2013.

- Mechanistic models

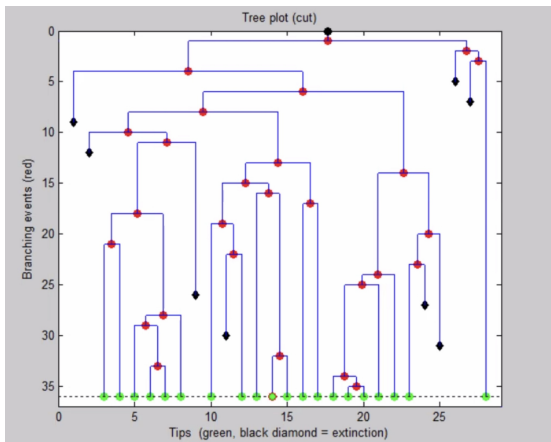


Figure: Simulation under some assumptions about the branching and extinction processes.

● Species tree: Desired features

- Evolution of a branch depends on underlying phases.
- Initial distribution of phases on a branch may depend on the parent branch.
- After birth, a branch evolves independently of all other branches.
- Possible events: speciation, extinction, phase transition.

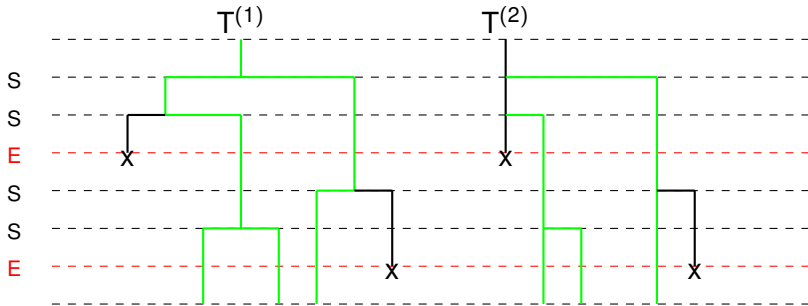
Therefore, we apply Markovian Binary Tree (MBT) model. ⁴

⁴Nectaros Kontoleon. The Markovian binary tree : A model of the macroevolutionary process. PhD thesis, The University of Adelaide, 2006.

● A wide range of speciation types

- I. Each of two new branches starts as a new process;
- II. Each of two new branches is a continuation of the parent branch;
- III. The left branch is the continuation of the parent branch while the right branch is the new species;

... etc



● Markovian Binary Tree (MBT) model

MBT $\{(M(t), \varphi(t)) : t \geq 0\}$ with state space

$$\mathcal{S} = \mathbb{N} \times \{E, 1, \dots, n\}$$

where

- level variable $M(t) \in \mathbb{N}$ counts the number of species born in $[0, t]$
- phase variable $\varphi(t) \in \{E, 1, \dots, n\}$ records some information that drives the evolution of the species
- $\{\varphi(t) : t \geq 0\}$ is a continuous-time Markov chain with an absorbing state E (extinction).

● MBT model parameters

$$\boldsymbol{\alpha}^{(0)} = [\alpha_j^{(0)}]_{j=1,\dots,n} \quad \mathbf{d} = [d_i]_{i=1,\dots,n} \quad \mathbf{D}_0 = [(D_0)_{ij}]_{i,j=1,\dots,n}$$

$$\mathbf{D}_1 = [(D_1)_{ij}]_{i,j=1,\dots,n} \quad \mathbf{P} = [P_{j,ik}]_{i,j,k=1,\dots,n} \quad \mathbf{B} = [B_{i,jk}]_{i,j,k=1,\dots,n}$$

- $\alpha_j^{(0)}$ probability that a branch starts in phase j .
- d_i rate at which a branch terminates when in phase i .
- $(D_0)_{ij}$ rate at which a branch changes phase from i to j .
- $(D_1)_{ij}$ rate at which a branch gives birth and simultaneously transitions to phase j when in phase i .
- $P_{j,ik}$ probability that a branch starts in phase j given its parent transitioned from phase i to phase k at the time of giving birth.
- $B_{i,jk} = (D_1)_{ik} P_{j,ik}$ rate at which a parent in phase i makes a transition to phase k and simultaneously gives birth to child in phase j .

Example¹: BiSSE model⁵ as an MBT

$$\begin{aligned}
 \mathcal{S} &= \mathbb{N} \times \{E, 0, 1\}, \quad \boldsymbol{\alpha}^{(0)} = \begin{bmatrix} \alpha_0^{(0)} & \alpha_1^{(0)} \end{bmatrix}, \\
 \mathbf{d} &= \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \quad \mathbf{D}_0 = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}, \quad \mathbf{D}_1 = \begin{bmatrix} \lambda_0 & 0 \\ 0 & \lambda_1 \end{bmatrix}, \\
 \mathbf{P} &= \begin{bmatrix} P_{0,00} & P_{1,00} \\ P_{0,01} & P_{1,01} \\ P_{0,10} & P_{1,10} \\ P_{0,11} & P_{1,11} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \\
 \mathbf{B} &= \begin{bmatrix} B_{0,00} & B_{0,01} & B_{1,00} & B_{1,01} \\ B_{0,10} & B_{0,11} & B_{1,10} & B_{1,11} \end{bmatrix} = \begin{bmatrix} \lambda_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}.
 \end{aligned}$$

⁵W.P. Maddison, P.E. Midford, and S.P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, 2007.

Example¹: MuSSE model⁶ as an MBT

$$\begin{aligned}
 \mathcal{S} &= \mathbb{N} \times \{E, 1, \dots, n\} \\
 \alpha^{(0)} &= [\alpha^{(0)}]_{i=1,2,\dots,n} \\
 \mathbf{d} &= ([\mu_i]_{i=1,2,\dots,n})^T \quad (\text{column vector}) \\
 \mathbf{D}_0 &= [q_{ij}]_{i=1,2,\dots,n} \\
 \mathbf{D}_1 &= \text{diag}(\lambda_i)_{i=1,2,\dots,n} \\
 \mathbf{P} &= [P_{j,ik}]_{i,j,k=1,\dots,n} \text{ is such that} \\
 &\quad P_{j,ik} = 1 \text{ for } i = j = k \text{ and } P_{j,ik} = 0 \text{ otherwise} \\
 \mathbf{B} &= [B_{i,jk}]_{i,j,k=1,\dots,n} \text{ is such that} \\
 &\quad B_{i,jk} = \lambda_i \text{ for } i = j = k \text{ and } B_{i,jk} = 0 \text{ otherwise.}
 \end{aligned}$$

⁶R.G. FitzJohn. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3(6):1084–1092, 2012.

MBT is

- a particular class of a level-dependent Quasi-Birth-and-Death-Process (QBD) ^{7 8 9}

and

- a particular class of continuous-time multi-type branching processes ¹⁰ in which the life of each branch is a Markovian arrival process (MAP) ¹¹.

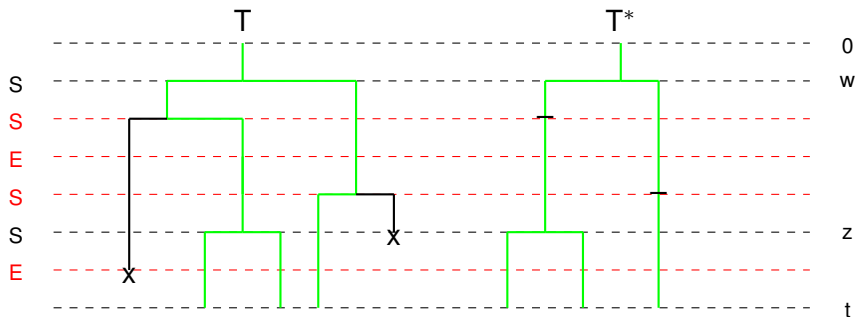
⁷Neuts MF (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. The Johns Hopkins University Press, Baltimore.

⁸Neuts MF (1989) Structured stochastic matrices of M/G/1 type and their applications. Marcel Dekker, New York.

⁹T. Phung-Duc, H. Masuyama, S. Kasahara, and Y. Takahashi. A simple algorithm for the rate matrices of level-dependent QBD processes. In Proceedings of the 5th International Conference on Queueing Theory and Network Applications, pages 46–52, 2010.

¹⁰K. Athreya and P. Ney. Branching Processes. Springer-Verlag, New York, 1972.

¹¹G. Latouche, M.-A. Remiche, and P. Taylor. Transient Markov arrival processes. The Annals of Applied Probability, 13(2):628–640, 2003.

True tree T versus reconstructed species tree T^* 

- $[\mathbf{G}(w, z; t)]_{ij}$ likelihood of observing the internal branch (w, z)
- $[\mathbf{D}^{(1)}(t - w)]_{k\ell}$ likelihood of observing the external branch (w, t)

$$\mathbf{G}(0, t; t) \equiv \mathbf{D}^{(1)}(t)$$

Key derived measures

- Likelihood

$$\ell(T^*)$$

of observing reconstructed tree T^* .

- (time-dependent) Probability

$$D_{i|n}(t)$$

that a tree with n tips at time t has i tips on the left subtree.

- Long-run probability

$$D_{i|n} = \lim_{t \rightarrow \infty} D_{i|n}(t)$$

that a tree with n tips has i tips on the left subtree.

Results summary

We derive expressions for $\ell(T^*)$ under general MBT model.

As (a simpler) example, under variant I¹² of the model, we have

$$\ell(T^*) = \prod_{k=1}^K \left(\alpha \mathbf{G}(t_0^k, t_1^k; t) \mathbf{D}_1 \mathbf{1} \right) \prod_{m=1}^M \left(\alpha \mathbf{D}^{(1)}(t - \hat{t}_0^m) \right) \quad (1)$$

- $\alpha \mathbf{G}(t_0^k, t_1^k; t) \mathbf{D}_1 \mathbf{1}$ is the likelihood of observing the internal branch (t_0^k, t_1^k)
- $\alpha \mathbf{D}^{(1)}(t - \hat{t}_0^m)$ is the likelihood of observing the external branch (\hat{t}_0^m, t) .

¹²A. Ch. Soewongsono, B.R. Holland and M.M. O'Reilly. Tree Shape Statistics of Trees Generated Using Phase-Type Distributed Times to Speciation. *To be submitted.*

Results summary

We derive results for $\ell(T^*)$ via differential equations and iterations.

These results involve

- the probability $D^{(i,n)}(t)$ that we see i tips on the left and n total tips at time t , and
- the probability $D^{(n)}(t)$ that we see n tips at the time t .

Therefore, we obtain the result for

$$D_{i|n}(t) = \frac{D^{(i,n)}(t)}{D^{(n)}(t)} \quad (2)$$

which can be computed for a given set of parameters of the model.

Some DEs

Well known ¹³ (+ solution methods):

$$\begin{aligned}\frac{d\mathbf{E}(t)}{dt} &= \mathbf{d} + \mathbf{D}_0\mathbf{E}(t) + \mathbf{B}(\mathbf{E}(t) \otimes \mathbf{E}(t)) \\ \mathbf{E}(0) &= \mathbf{0}.\end{aligned}\quad (3)$$

One of our results¹ (+ solution methods):

$$\begin{aligned}\frac{d\mathbf{D}^{(n)}(t)}{dt} &= \mathbf{D}_0\mathbf{D}^{(n)}(t) + \sum_{i=0}^n \mathbf{B} \left(\mathbf{D}^{(i)}(t) \otimes \mathbf{D}^{(n-i)}(t) \right) \\ \mathbf{D}^{(n)}(0) &= I(n=1)\mathbf{1}.\end{aligned}\quad (4)$$

¹³Remiche M.-A., Hautphenne S., Latouche G. Transient features for Markovian binary trees. In Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, 2009.

Approximation of $D_{i|n}$ based on data

The probability $D_{i|n}$ that a tree with n tips has i tips on the left subtree can be estimated using statistical methods and the formula¹⁴

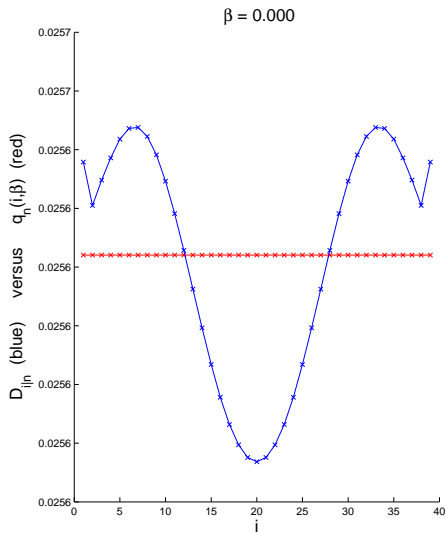
$$q_n(i, \beta) = \frac{1}{\alpha_n(\beta)} \frac{\Gamma(\beta + i + 1)\Gamma(\beta + n - i + 1)}{\Gamma(i + 1)\Gamma(n - i + 1)}, \quad 1 \leq i \leq n - 1 \quad (5)$$

where $\alpha_n(\beta)$ is a normalising constant.

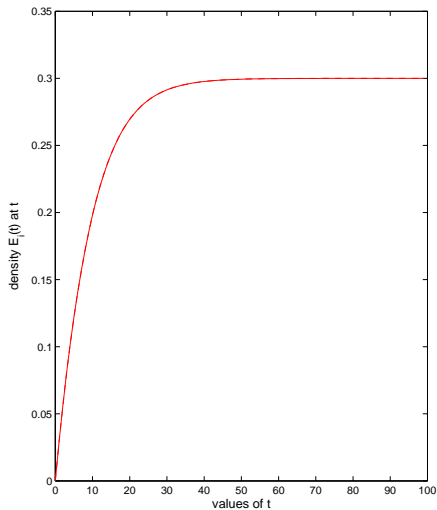
We compare this estimate with $D_{i|n}$.

¹⁴D.J. Aldous. 1996. Probability distributions on cladograms. Pages 1–18 in Random discrete structures (D. J. Aldous, and R. Pemantle, eds.). Springer, New York.

Example: $\mu_0 = \mu_1 = 0.03$, $\lambda_0 = \lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$ $E_0 = E_1 = 0.3$

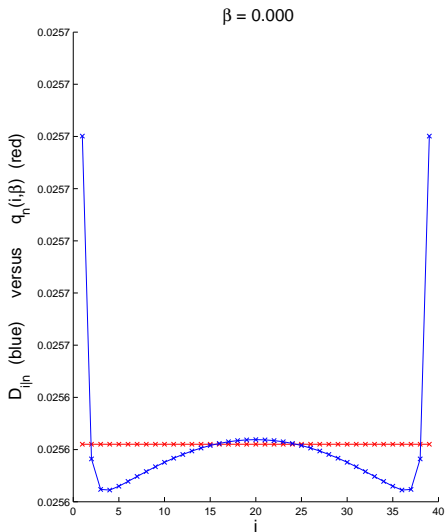


Example: $\mu_0 = \mu_1 = 0.03$, $\lambda_0 = \lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$ $E_0 = E_1 = 0.3$



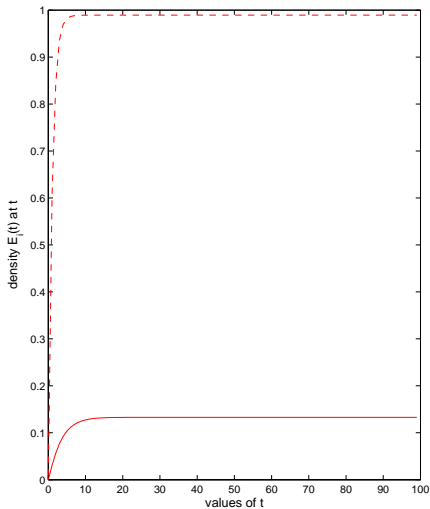
Example: $\mu_0 = 0.03$, $\mu_1 = 0.9$, $\lambda_0 = 0.3$, $\lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$

$E_0 = 0.1329$, $E_1 = 0.9893$



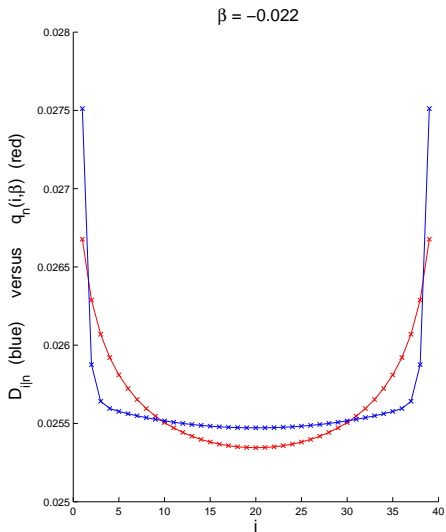
Example: $\mu_0 = 0.03$, $\mu_1 = 0.9$, $\lambda_0 = 0.3$, $\lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$

$E_0 = 0.1329$, $E_1 = 0.9893$



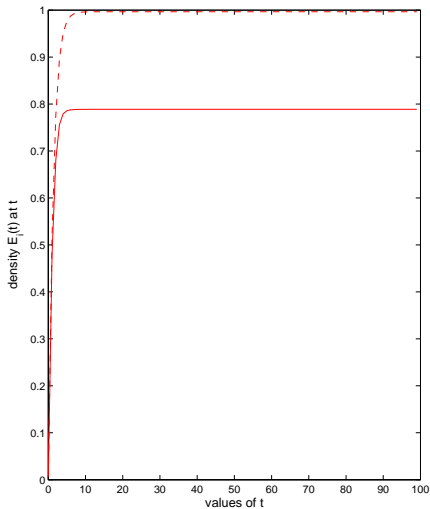
Example: $\mu_0 = \mu_1 = 0.7$, $\lambda_0 = 0.9$, $\lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$

$E_0 = 0.7887$, $E_1 = 0.9965$



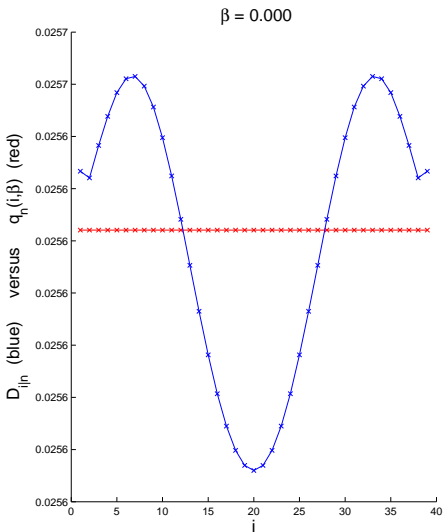
Example: $\mu_0 = \mu_1 = 0.7$, $\lambda_0 = 0.9$, $\lambda_1 = 0.1$, $q_{01} = q_{10} = 0.01$

$E_0 = 0.7887$, $E_1 = 0.9965$



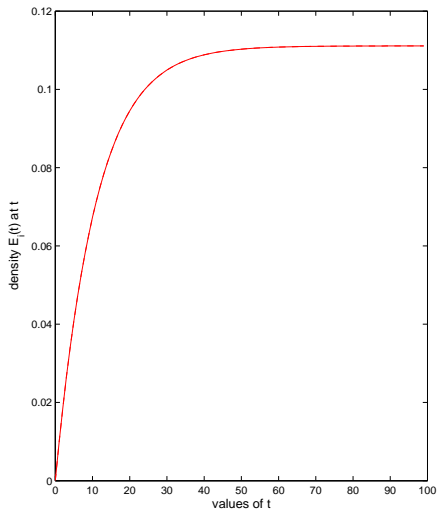
Example: $\mu_0 = \mu_1 = 0.03$, $\lambda_0 = \lambda_1 = 0.09$, $q_{01} = 0.005$, $q_{10} = 0.010$

$E_0 = E_1 = 0.1111$



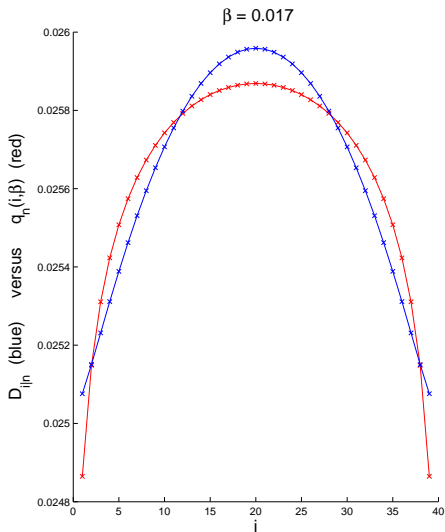
Example: $\mu_0 = \mu_1 = 0.03$, $\lambda_0 = \lambda_1 = 0.09$, $q_{01} = 0.005$, $q_{10} = 0.010$

$$E_0 = E_1 = 0.1111$$



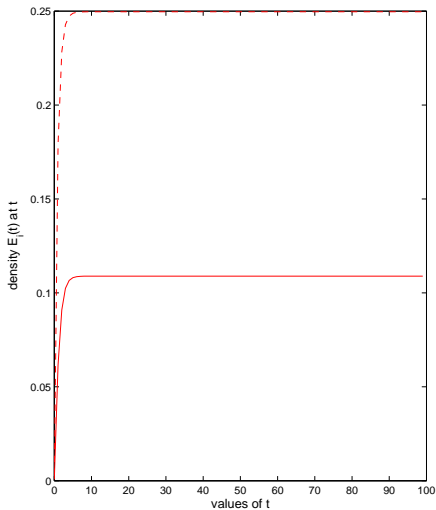
Example: $\mu_0 = 0.0756$, $\mu_1 = 0.3138$, $\lambda_0 = 0.9460$, $\lambda_1 = 0.8545$, $q_{01} = 0.1735$,

$q_{10} = 0.5359$ $E_0 = 0.1089$, $E_1 = 0.2496$



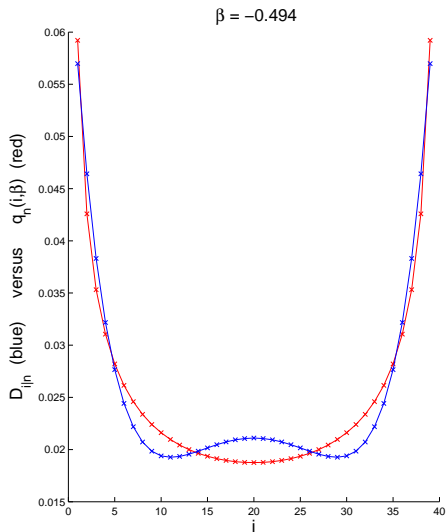
Example: $\mu_0 = 0.0756$, $\mu_1 = 0.3138$, $\lambda_0 = 0.9460$, $\lambda_1 = 0.8545$, $q_{01} = 0.1735$,

$q_{10} = 0.5359$ $E_0 = 0.1089$, $E_1 = 0.2496$



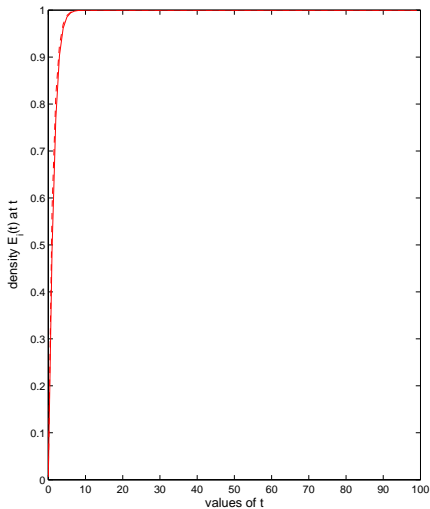
Example: $\mu_0 = 0.5840$, $\mu_1 = 0.9091$, $\lambda_0 = 0.2896$, $\lambda_1 = 0.3542$, $q_{01} = 0.3298$,

$q_{10} = 0.7271$ $E_0 = E_1 = 1$



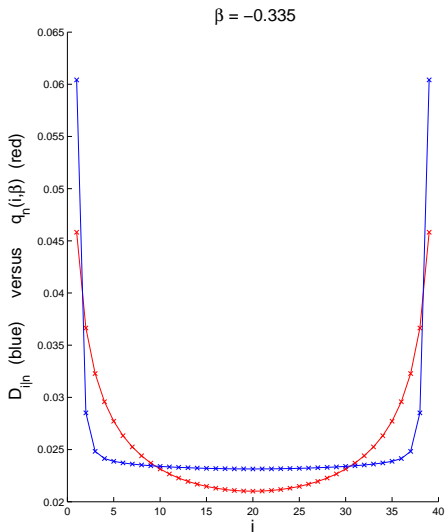
Example: $\mu_0 = 0.5840$, $\mu_1 = 0.9091$, $\lambda_0 = 0.2896$, $\lambda_1 = 0.3542$, $q_{01} = 0.3298$,

$q_{10} = 0.7271$ $E_0 = E_1 = 1$



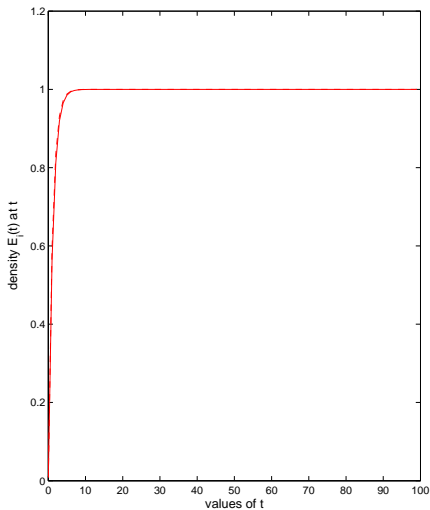
Example: $\mu_0 = 0.8140$, $\mu_1 = 0.8874$, $\lambda_0 = 0.0120$, $\lambda_1 = 0.2498$, $q_{01} = 0.0534$,

$q_{10} = 0.3518$ $E_0 = E_1 = 1$

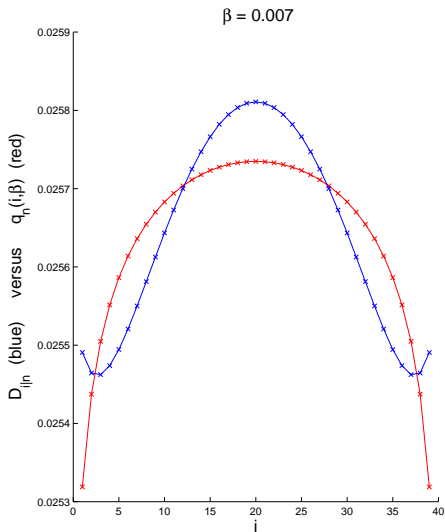


Example: $\mu_0 = 0.8140$, $\mu_1 = 0.8874$, $\lambda_0 = 0.0120$, $\lambda_1 = 0.2498$, $q_{01} = 0.0534$,

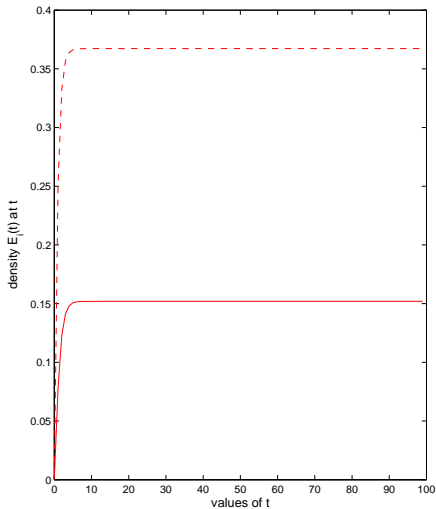
$q_{10} = 0.3518$ $E_0 = E_1 = 1$



Example: $\mu_0 = 0.0796$, $\mu_1 = 0.4009$, $\lambda_0 = 0.9000$, $\lambda_1 = 0.8747$, $q_{01} = 0.2252$,
 $q_{10} = 0.2343$, $E_0 = 0.1519$, $E_1 = 0.3672$



Example: $\mu_0 = 0.0796$, $\mu_1 = 0.4009$, $\lambda_0 = 0.9000$, $\lambda_1 = 0.8747$, $q_{01} = 0.2252$,
 $q_{10} = 0.2343$, $E_0 = 0.1519$, $E_1 = 0.3672$



● Gene tree: Desired features

- Evolution of each branch may depend on other branches, due to interactions between the genes which share various functions.
- No extinction of the gene family due to the protective mechanisms, and so we model the evolution of a gene family that has survived.
- Evolution of a branch depends on underlying phases.
- Possible events: gene duplication, gene loss, neofunctionalisation, nonfunctionalisation.

Therefore, we apply Quasi-Birth-and-Death (QBD) model ¹⁵.

¹⁵J. Diao, T.L. Stark, D.A. Liberles, M.M. O'Reilly, and B.R. Holland. Level-dependent QBD models for the evolution of a family of gene duplicates. *Stochastic Models*, 36(2):285–311, 2020.

- Quasi-Birth-and-Death (QBD) model

QBD $\{(Y(t), \varphi(t)) : t \geq 0\}$ with state space

$$\mathcal{S} = \{(n, k) : n = 1, 2, \dots; k = 1, \dots, K_n\}$$

where

- level variable $Y(t)$ records the number of genes in the family
- phase variable $\varphi(t)$ records some information about the family.

- QBD model parameters

Initial distribution vector

$$\alpha_n = [\alpha_{n,k}]_{k=1,\dots,K_n}, \quad \alpha_{n,k} = \mathcal{P}(Y(0) = n, \varphi(0) = k)$$

and generator

$$\mathbf{Q} = [\mathbf{Q}^{[n,n']}]_{n,n'} = \begin{bmatrix} \mathbf{Q}^{[1,1]} & \mathbf{Q}^{[1,2]} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{Q}^{[2,1]} & \mathbf{Q}^{[2,2]} & \mathbf{Q}^{[2,3]} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{Q}^{[3,2]} & \mathbf{Q}^{[3,3]} & \mathbf{Q}^{[3,4]} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}^{[4,3]} & \mathbf{Q}^{[4,4]} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where block $\mathbf{Q}^{[n,n']} = [q_{(n,k)(n',k')}]$ records the transition rates $q_{(n,k)(n',k')}$ from states (n, k) to (n', k') , $n' \in \{n-1, n, n+1\}$.

● Example¹:

Consider a QBD with $\{(Y(t), \varphi(t)) : t \geq 0\}$ with state space

$$\mathcal{S} = \{(n, m, n_{(L)}, m_{(L)}, k)\}$$

where

- $n = 1, 2, \dots$ is the number of genes of family
- $m = 0, \dots, n$ is the number of redundant genes
- $n_{(L)} = 1, \dots, n - 1$ is the number of genes on the left branch
- $m_{(L)} = 0, \dots, m$ is the number of the redundant genes on the left branch
- $k = 1, \dots, K$ is some information about the gene family used to model the total number of functions in the family.

We derive generator of such QBD in terms of parameters u_c, u_d, u_r and u_f , by modifying the results in Table 1 in¹⁵.

Example cont: parameters and stability condition

Stability condition: ¹⁶ π exists $\iff u_c + u_r > u_d$ due to

$$u_c + u_r > u_d \implies \lambda_{\text{Lost}} = (u_r + u_c) \cdot m > \lambda_{\text{Dup}} = u_d \cdot n$$

$$\text{for all } n > \frac{u_r + u_c}{u_r + u_c - u_d} \cdot z_0$$

- u_c per gene rate of null mutation in the coding region of a gene;
- u_r per region rate of null mutation in regulatory region of a gene;
- u_d per gene rate of duplication of a gene;
- u_f per gene rate of obtaining a new function of a gene;
- z_0 the number of functions in the gene family.

¹⁶J. Diao, B.R. Holland and M.M. O'Reilly. A subfunctionalization model of gene family evolution predicts balanced tree shapes. *To be submitted*.

- Key derived measures (transient and long-run)

- Likelihood

$$\ell(T^*)$$

of observing reconstructed tree T^* .

- Probability

$$p(i|n)(t)$$

that a tree with n tips at time t has i tips on the left subtree.

- Long-run probability

$$p(i|n) = \lim_{t \rightarrow \infty} p(i|n)(t)$$

that a tree with n tips has i tips on the left subtree.

● Results summary

- We state expressions for these measures via Laplace transforms and iterations using the theory of QBDs.
- We outline an algorithm for the computation of these measures for any set of parameters of the model.

For example,


$$\rho(i|n) = \frac{\rho(i, n)}{\rho(n)}, \quad i = 1, \dots, n-1$$

$$\rho(n) = \sum_{i=1}^{n-1} \rho(i, n)$$

$$\rho(i, n) = \sum_{m, m_{(L)}, k} \pi_{(n, m, i, m_{(L)}, k)}$$

where $\pi = [\pi_{(n, m, n_{(L)}, m_{(L)}, k)}]$ is the stationary distribution.

Current and future work

- Computations and code¹:
 - ▶ Simulations
 - ▶ Computations for the species tree model (**generalized** MBT) 
 - ▶ Computations for the gene tree model
 - ▶ Computations for the reconciliation problem:
(Collection of) best fitting gene trees
- Future work:
 - ▶ Applications of our results to large data sets.
 - ▶ Modelling extensions to include other evolutionary behaviours.



Thank you for listening

