Inferring population histories in the IM-model

Lars Nørvang Andersen Thomas Mailund Asger Hobolth

BiRC
Aarhus University

Montpellier, June 2012

BiRC

Nature, March 8

Scally et al. *Nature*, 483 169–175, Mar 2012.

Gor

The case of 3 lineages

Our data points are the alignment columns:

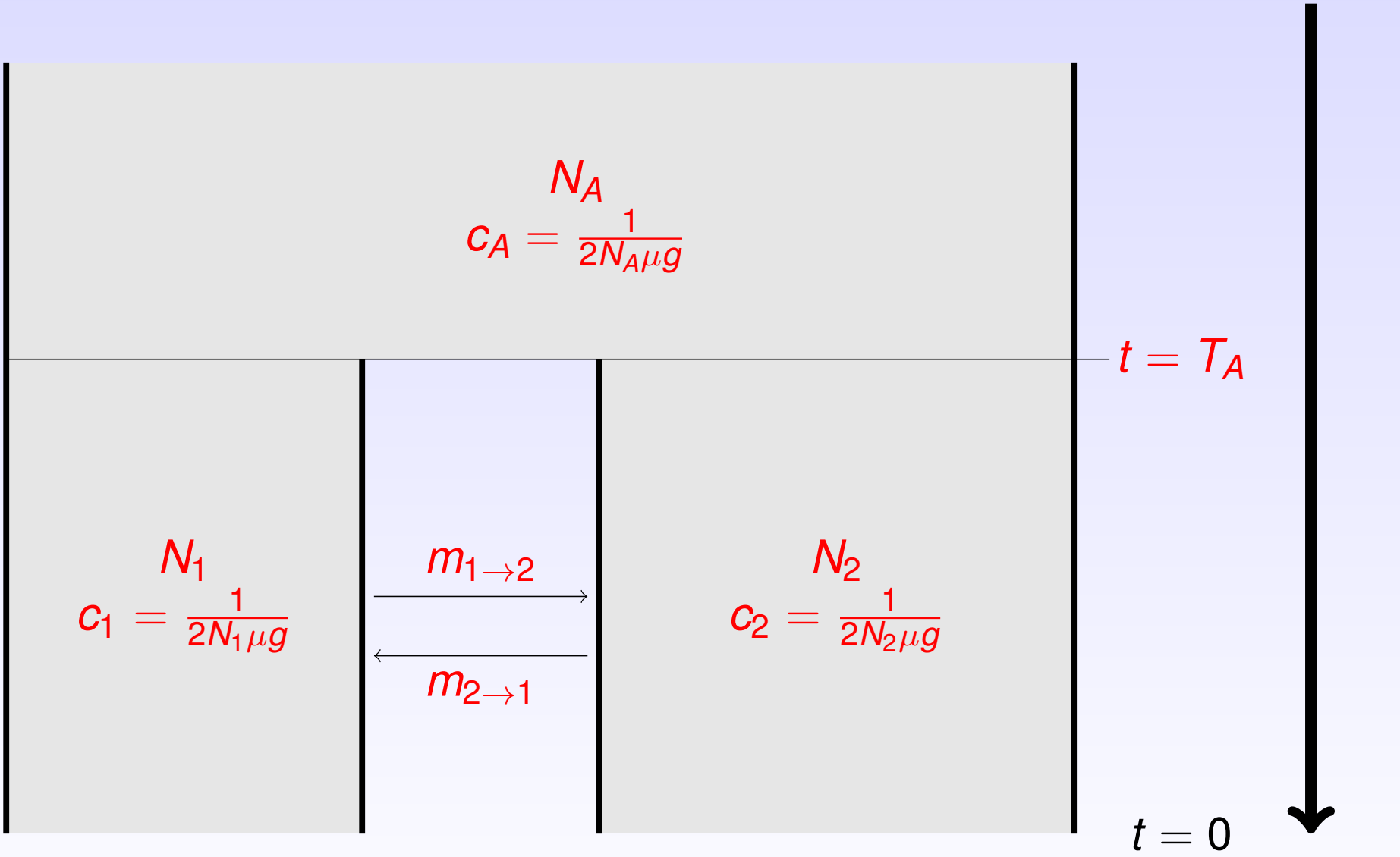
A A A	A A C	A A G	... ...	T T C	T T T
N_{AAA}	N_{AAC}	N_{AAG}	...	N_{TTC}	N_{TTT}

We want to analyze this data using maximum likelihood
Free recombination between loci implies that the log-likelihood function given the data **x** is given by

$$L(\Theta \mid \mathbf{x}) = \sum_{x \in \{A,C,T,G\}^3} N(x) \log \mathbb{P}(x)$$

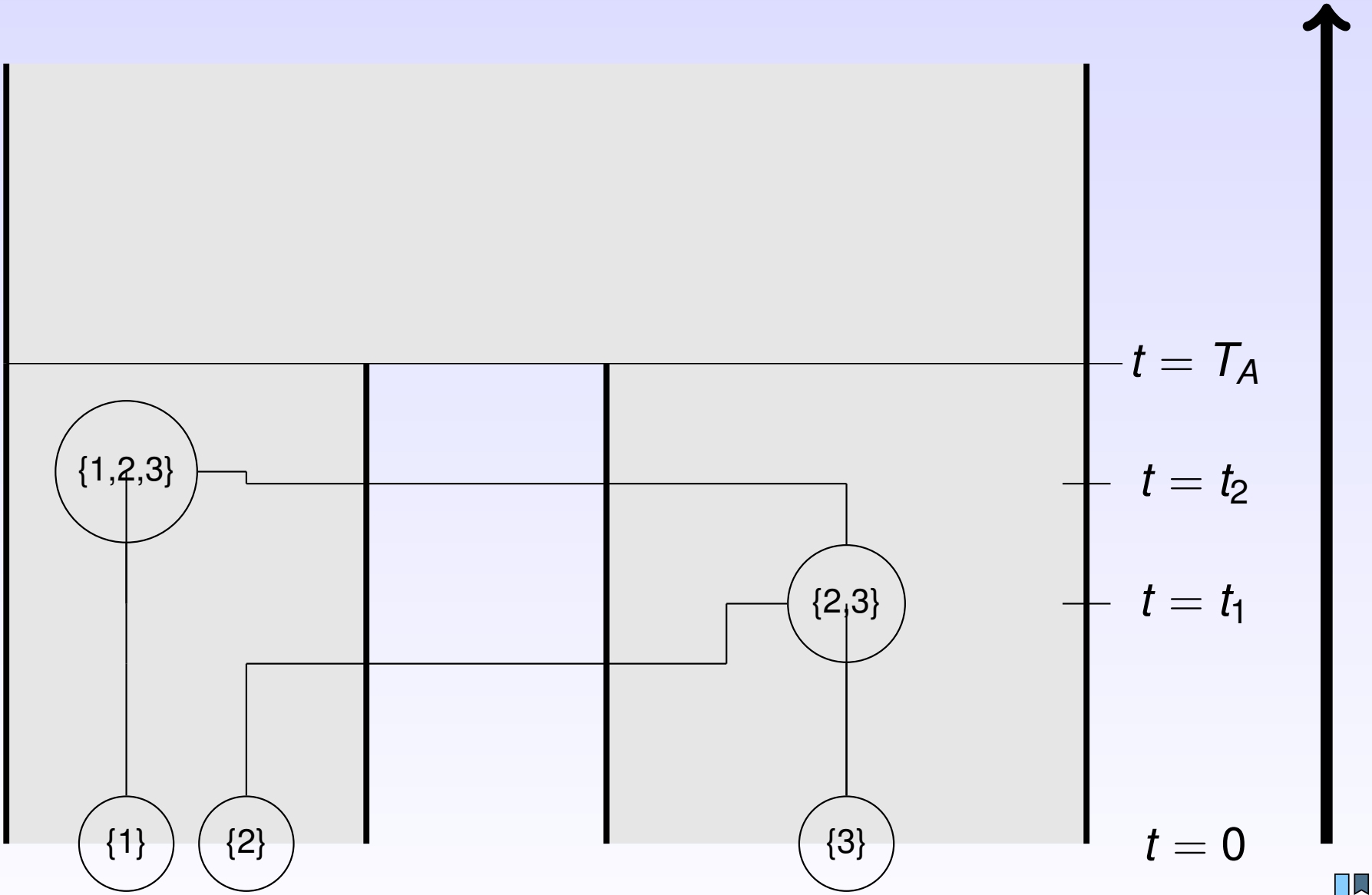
BiRC

The IM-model



BiRC

The IM-model



The IM-model - prior to T_A

The lineages are subsets $\ell_i \subseteq \{1, 2, \ldots, L\}$
The populations are number $j_i \in \{1, 2, \ldots, N\}$
Each state is a union of tuples: $\{(j_i, l_i) \mid i = 1, \ldots, m\}$
Examples of states:
Starting state: $s = \{(1, \{1\}), (1, \{2\}), (2, \{3\})\}$
Migration: $s_1 = \{(1, \{1\}), (2, \{2\}), (2, \{3\})\}$
Coalescence: $s_2 = \{(1, \{1\}), (2, \{2, 3\})\}$
Migration: $s_3 = \{(1, \{1\}), (1, \{2, 3\})\}$
Coalescence: $s_4 = \{(1, \{1, 2, 3\})\}$

BiRC

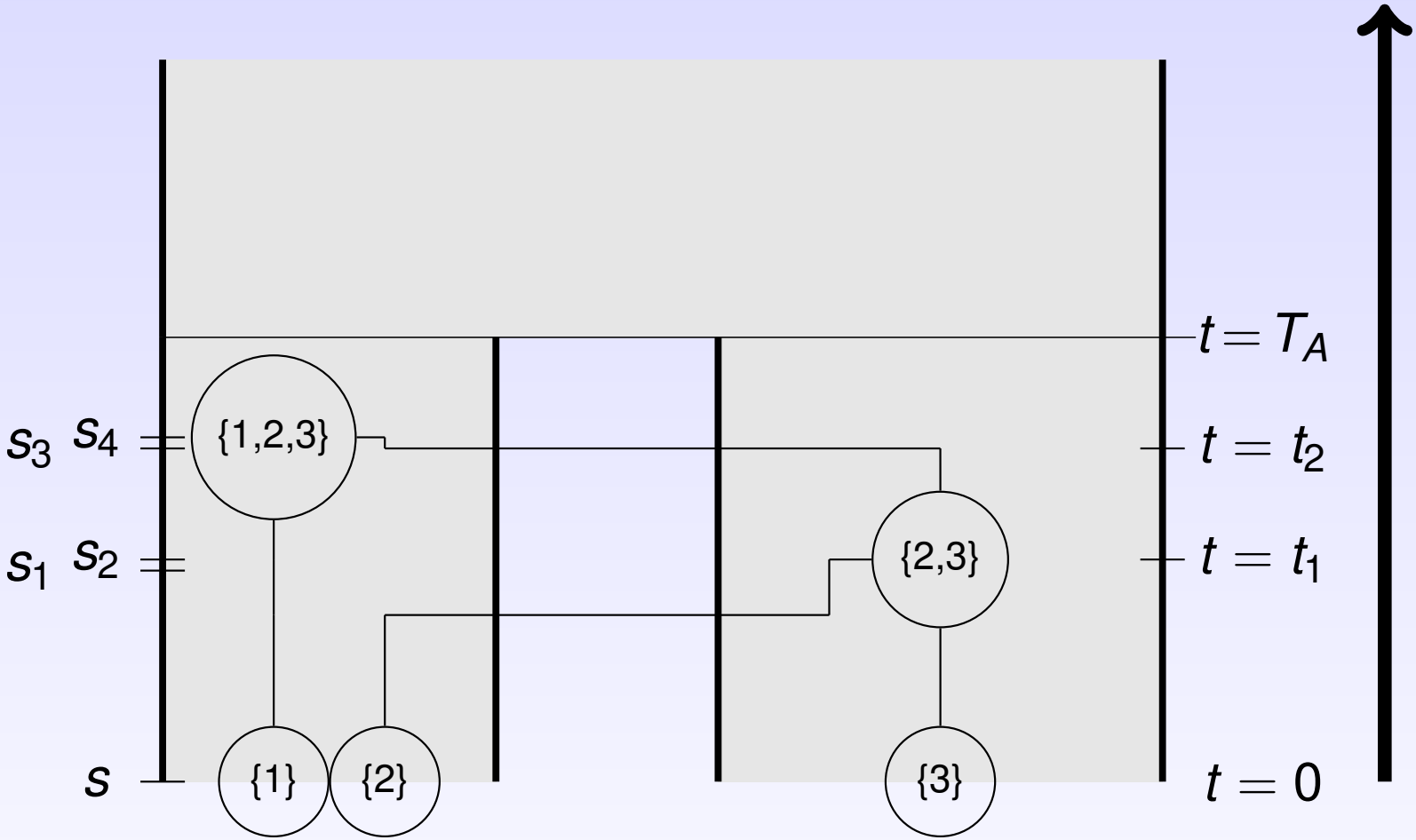
The IM-model - prior to T_A

$$Q_{\alpha,\beta} = \begin{cases} m_{i\to j} & \text{if } \alpha = \mathcal{S} \cup \{(i,l)\}\,,\ \beta = \mathcal{S} \cup \{(j,l)\} \\ c_i & \text{if } \alpha = \mathcal{S} \cup \{(i,l_1)\} \cup \{(i,l_2)\}\,,\ \beta = \mathcal{S} \cup \{(i,l_1 \cup l_2)\} \\ 0 & \text{otherwise}\,, \end{cases}$$

where $\mathcal{S}$ is of the form $\cup_i \{(j_i, l_i)\}$

BiRC

The IM-model



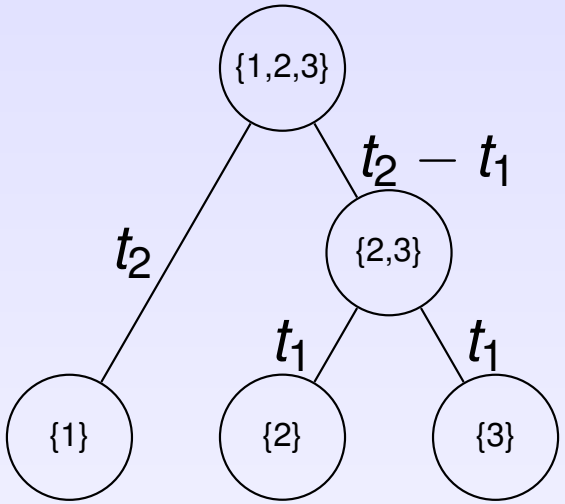
$s = \{(1, \{1\}), (1, \{2\}), (2, \{3\})\}$, $s_1 = \{(1, \{1\}), (2, \{2\}), (2, \{3\})\}$
$s_2 = \{(1, \{1\}), (2, \{2, 3\})\}$, $s_3 = \{(1, \{1\}), (1, \{2, 3\})\}$,$s_4 = \{(1, \{1, 2, 3\})\}$

$$\left(e^{Qt_1}\right)_{s,s_1} c_2 \left(e^{Q(t_2-t_1)}\right)_{s_2,s_3} c_1$$

BiRC

Genealogies - Coalescent trees

Sample paths of $X(t)$ are *genealogies*.
coalescent trees are unions of genealogies



$$f(\mathcal{C}(t_1, t_2)) = \sum_{(\alpha_i)} \left(e^{Qt_1}\right)_{s,s_1} c_1 \qquad \left(e^{Q(t_2-t_1)}\right)_{s_2,s_3} c_2$$

$$f(\mathcal{C}(t_1, t_2)) = \sum_{(\alpha_i)} \left(e^{Qt_1}\right)_{s,\alpha_1} c_{(\alpha_1,\alpha_2)} \left(e^{Q(t_2-t_1)}\right)_{\alpha_2,\alpha_3} c_{(\alpha_3,\alpha_4)}$$

BiRC

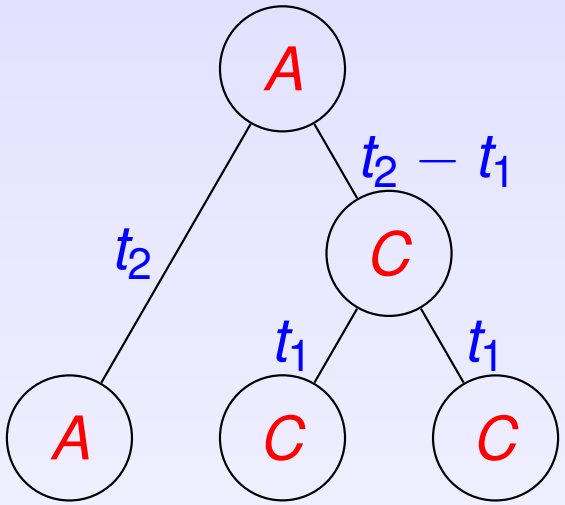
The nucleotide substitution matrix - time-reversible case

$$\overline{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_G & \gamma\pi_T \\ \alpha\pi_A & \cdot & \rho\pi_G & \sigma\pi_T \\ \beta\pi_A & \rho\pi_C & \cdot & \tau\pi_T \\ \gamma\pi_A & \sigma\pi_C & \tau\pi_G & \cdot \end{pmatrix} \end{array}$$

Stationary distribution: $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$

BiRC

Nucleotide assignment



$$\mathbb{P}(x \mid \mathcal{C}(t_1, t_2)) = \pi_A \left(e^{\bar{Q}t_2}\right)_{A,A} \left(e^{\bar{Q}(t_2-t_1)}\right)_{A,C} \left(e^{\bar{Q}t_1}\right)_{C,C} \left(e^{\bar{Q}t_1}\right)_{C,C}$$

BiRC

Excplicit calculations

Recall:

$$f(\mathcal{C}(t_1,t_2)) = \sum_{(\alpha_i)} \left(\mathrm{e}^{Qt_1}\right)_{s,\alpha_1} c_{(\alpha_1,\alpha_2)} \left(\mathrm{e}^{Q(t_2-t_1)}\right)_{\alpha_2,\alpha_3} c_{(\alpha_3,\alpha_4)}$$

$$\begin{aligned}
\mathbb{P}(x) &= \int\int \mathbb{P}(x \mid \mathcal{C}(t_1,t_2) f(\mathcal{C}(t_1,t_2)) \mathrm{d}t_1 \mathrm{d}t_2 \\
&= \sum C \int\int \left(\mathrm{e}^{\bar{Q}t_2}\right)_{A,A} \ldots \left(\mathrm{e}^{Q(t_2-t_1)}\right)_{\alpha_2,\alpha_3} \mathrm{d}t_1 \mathrm{d}t_2 \\
&= \sum\sum CC_1 \int\int \mathrm{e}^{\bar{\lambda}_i t_2} \ldots \mathrm{e}^{\lambda_j (t_2-t_1)} \mathrm{d}t_1 \mathrm{d}t_2
\end{aligned}$$

$$\left(\mathrm{e}^{\bar{Q}t}\right)_{x,y} = \sum_i \bar{v}_{x,i} \bar{v}_{i,y}^{-1} \mathrm{e}^{\bar{\lambda}_i t},$$

BiRC

Statespace explosion

The number of states increases very rapidly:

L	2	3	4	5	6
States	6	22	94	454	2430

Table : The number of states for $L = 2, 3, 4, 5, 6$ lineages

BiRC

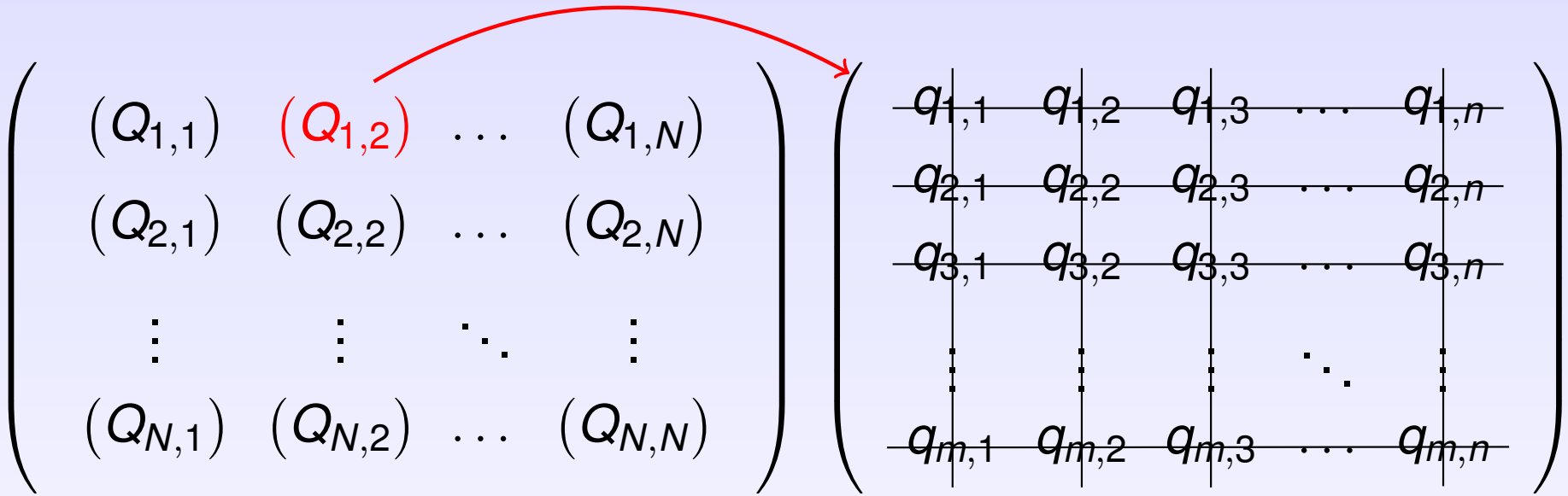
Lumpability

Given CTMC $\{X_t\}_{t\geq 0}$,(S, Q, π_0), S state space, Q rate matrix, π_0 inititial distribution.
Partition $\mathcal{P} = \{S_i\}_{i\in I}$
Define a stochastic process $\hat{X}_n = i \leftrightarrow X_n \in S_i$.
Under what conditions $\{\hat{X}_n\}$ a CTMC?

BiRC

Strong Lumpability - Exact Lumpability

$$\mathcal{P} = \{\{S_1\}, \{S_2\}, \ldots, \{S_N\}\}$$



- Strong Lumpability : $\forall I \neq J, i, i' \in S_I : \sum_{j \in J} q_{i,j} = \sum_{j \in J} q_{i',j}$
- Exact Lumpability :
  $\forall S_I, S_J, j, j' \in S_J : \sum_{i \in S_I} q_{i,j} = \sum_{i \in S_I} q_{i,j'}$

BiRC

Applying lumpability

Q :

	1 2 3 |	1 2 | 3	1 3 | 2	2 3 | 1	1 | 2 3	2 | 1 3	3 | 1 2	| 1 2 3	A
1 2 3 |	$\cdot$	$m_{1\to 2}$	$m_{1\to 2}$	$m_{1\to 2}$	0	0	0	0	$3c_1$
1 2 | 3	$m_{2\to 1}$	$\cdot$	0	0	$m_{1\to 2}$	$m_{1\to 2}$	0	0	c_1
1 3 | 2	$m_{2\to 1}$	0	$\cdot$	0	$m_{1\to 2}$	0	$m_{1\to 2}$	0	c_1
2 3 | 1	$m_{2\to 1}$	0	0	$\cdot$	0	$m_{1\to 2}$	$m_{1\to 2}$	0	c_1
1 | 2 3	0	$m_{2\to 1}$	$m_{2\to 1}$	0	$\cdot$	0	0	$m_{1\to 2}$	c_2
2 | 1 3	0	$m_{2\to 1}$	0	$m_{2\to 1}$	0	$\cdot$	0	$m_{1\to 2}$	c_2
3 | 1 2	0	0	$m_{2\to 1}$	$m_{2\to 1}$	0	0	$\cdot$	$m_{1\to 2}$	c_2
| 1 2 3	0	0	0	0	$m_{2\to 1}$	$m_{2\to 1}$	$m_{2\to 1}$	$\cdot$	$3c_2$
A	0	0	0	0	0	0	0	0	0

We wish to calculate:

$$\mathbb{P}(X_t = (1 \mid 2\ 3) \mid X_0 = (1\ 2\ 3 \mid)) = \mathbb{P}_\pi(X_t = 1 \mid 2\ 3) = \left(\mathrm{e}^{Qt}\right)_{1\ 2\ 3\ \mid, 1 \mid 2\ 3}$$

$\tilde{f} : 1 \mapsto 0, 2 \mapsto 0, 3 \mapsto 0$

BiRC

Applying lumpability

$\hat{Q}$:

$$\begin{array}{c} \\ 0\,0\,0\,| \\ 0\,0\,|\,0 \\ 0\,|\,0\,0 \\ |\,0\,0\,0 \\ A \end{array} \begin{array}{c} \begin{array}{ccccc} 0\,0\,0\,| & 0\,0\,|\,0 & 0\,|\,0\,0 & |\,0\,0\,0 & A \end{array} \\ \begin{pmatrix} \cdot & 3m_{1\to 2} & 0 & 0 & 3c_1 \\ m_{2\to 1} & \cdot & 2m_{1\to 2} & 0 & c_1 \\ 0 & 2m_{2\to 1} & \cdot & 0 & c_2 \\ 0 & 0 & 2m_{2\to 1} & \cdot & 3c_2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\mathbb{P}_{\pi}(X_t = (1 \mid 2\,3)) = \frac{1}{3}\mathbb{P}_{\hat{\pi}}(\hat{X}_t = (0 \mid 0\,0)) = \frac{1}{3}\left(\mathrm{e}^{\hat{Q}t}\right)_{0\,0\,0\,|,0\,|\,0\,0}$$

Reduced statespace

L	2	3	4	5	6
States	6	22	94	454	2430

Table : The number of states for $L = 2, 3, 4, 5, 6$ lineages

L	2	3	4	5	6
States	7	17	39	67	117

Table : The number of states in the reduced states-space for $L = 2, 3, 4, 5, 6$ lineages.

BiRC

Thank you.