Monophyletic concordance between species trees and gene genealogies with multiple mergers

Bjarki Eldon and James Degnan
Phylomania 2010
University of Tasmania
November 4-5, 2010
Low offspring number models

Kingman (1982) introduced the \( n \)-coalescent from an exchangeable Cannings offspring model; let \( \nu_i \) denote the number of offspring of individual \( i \)

\[
\mathbb{E}[^k_1] < \infty \quad \text{as} \quad N \to \infty; \quad k \geq 1
\]

Möhle and Sagitov (2001) characterised coalescent processes based on the timescale \( c_N \)

\[
c_N = \frac{\mathbb{E}[^1_1(\nu_1 - 1)]}{N - 1}
\]
Conditions for convergence to Kingman’s coalescent

Wright-Fisher and Moran models are exchangeable Cannings models with

\[ \lim_{N \to \infty} \frac{\mathbb{E}[\nu_1(\nu_1 - 1)(\nu_1 - 2)]}{N^2 c_N} = 0 \]

implying \( c_N \to 0 \) and convergence to Kingman’s coalescent.
High variance in offspring distribution

Ecology, reproductive biology, and genetics of a diverse group of marine organisms suggest many offspring contributed by few individuals (Beckenbach 94; Hedgecock 94)

Direct genotyping of parents and offspring provides evidence of large families in Pacific oyster (Boudry et al. 2002) and Lion-Paw Scallop (Petersen et al. 2008)

Cod, oysters, mussels, barnacles, sea stars, plants?
Evidence for large offspring distribution

- broadcast spawning and external fertilization
- high initial mortality
- very large population sizes
- low genetic diversity
- large number of singleton genetic variants
Donnelly and Kurtz (1999), Pitman (1999), and Sagitov (1999) independently introduce a multiple merger coalescent; $\Lambda$-coalescent with coalescence rate

$$
\lambda_{b,k} = \binom{b}{k} \int_0^1 x^k (1 - x)^{b-k} x^{-2} \Lambda(dx)
$$

Kingman’s coalescent is obtained if $\Lambda = \delta_0$

For simultaneous multiple merger coalescent processes, see Schweinsberg (2000) and Möhle and Sagitov (2001).
Schweinsberg’s heavy-tail model

Schweinsberg (2003)
Each individual produces a random number $X_i$ of potential offspring; $C > 0$ and $a > 0$ and constant population size $N$

$$\mathbb{P}[X_i \geq k] \sim \frac{C}{k^a}$$

and

$$\mathbb{E}[X_i] > 1$$

From the pool of potential offspring, sample without replacement to form the new generation
Coalescent process depends on $a$

Coalescent timescale in units of $c_N \sim N^{a-1}$ if $1 < a < 2$

<table>
<thead>
<tr>
<th>case</th>
<th>coalescent</th>
<th>coalescence rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \geq 2$</td>
<td>Kingman coalescent</td>
<td>$\binom{b}{2}$</td>
</tr>
<tr>
<td>$1 \leq a &lt; 2$</td>
<td>$\Lambda \sim \text{Beta}(2 - a, a)$</td>
<td>$\binom{b}{k} \frac{B(k - a, b - k + a)}{B(2 - a, a)}$</td>
</tr>
<tr>
<td>$0 &lt; a &lt; 1$</td>
<td>$\Xi$-coalescent</td>
<td></td>
</tr>
</tbody>
</table>
Eldon and Wakeley (2006)

A modified Moran model, in which the offspring number $U$ is random rather than fixed at one as in the usual Moran model

$$\mathbb{P}[U = u] = (1 - \varepsilon_N)\delta_2 + \varepsilon_N\delta_{[\psi N]}$$

and

$$\varepsilon_N \sim \frac{1}{N^\gamma}, \quad \gamma > 0$$
Coalescent process depends on $\gamma$

Coalescent timescale is $N_\gamma = \min \left( N^\gamma, N^2 \right)$, $\gamma > 0$

<table>
<thead>
<tr>
<th>case</th>
<th>coalescence rate</th>
<th>timescale</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma &gt; 2$</td>
<td>$\binom{n}{2}$</td>
<td>$N^2$</td>
</tr>
<tr>
<td>$\gamma = 2$</td>
<td>$\binom{b}{k} \left( \delta_2 + \psi^k(1 - \psi)^{b-k} \right)$</td>
<td>$N^2$</td>
</tr>
<tr>
<td>$\gamma &lt; 2$</td>
<td>$\binom{b}{k} \psi^k(1 - \psi)^{b-k}$</td>
<td>$N^\gamma$, $1 &lt; \gamma &lt; 2$</td>
</tr>
</tbody>
</table>
Ratios of coalescence times for $\Lambda = K + \Lambda_\psi$

○: $\overline{R}_1$; △: $\overline{R}_2$; ▽: $\overline{R}_3$; ◆: $\overline{R}_4$; +: $\overline{R}_{n-1}$
Ratios of coalescence times for $\Lambda = Beta(0.9, 1.1)$

- $\circ: R_1$
- $\triangle: R_2$
- $\triangledown: R_3$
- $\diamond: R_4$
- $+: R_{n-1}$
Ratios of coalescence times for \( \Lambda = Beta(0.1, 1.9) \)

- \( \circ : \overline{R_1} \)
- \( \triangle : \overline{R_2} \)
- \( \nabla : \overline{R_3} \)
- \( \diamond : \overline{R_4} \)
- \( + : \overline{R_{n-1}} \)

sample size \( n \)
Monophyletic concordance for Λ coalescents
Not monophyletic concordance
General form for $\mathbb{P}[MC]$ for two species

$$\mathbb{P}[MC] = \sum_{m_A, m_B} \mathbb{P}[MC; m_A, m_B] \mathbb{P}[m_A, m_B]$$

with

$$\mathbb{P}[n_A, n_B] = G_{n_A,m_A}(t)G_{n_B,m_B}(t)$$

and

$$\mathbb{P}[MC; m_A, m_B] = \sum_{k=2}^{m_A+m_B} \beta_{m_A+m_B,k} \left( \mathbb{P}[MC; m_A - k + 1, m_B] \binom{m_A}{k} \right)$$

$$+ \mathbb{P}[MC; m_A, m_B - k + 1] \binom{m_B}{k} / \binom{m_A + m_B}{k}$$
Computing $G_{i,j}(t)$

$G_{i,j}(t)$ is the probability of $j$ lines at time $t$ when starting from $i$ lines at time zero within one population.

A vector $c$ of ordered mergers associated with Kingman’s coalescent is simply $\{2, 2, \ldots, 2\}$

By way of example, starting from 10 lines, say, a coalescence sequence could be $\{3, 2, 5, 3\}$ in a $\Lambda$ coalescent.

Conditioning on the embedded chain, or the order of mergers.

Transition probabilities

\[
\beta_{i,j} = \begin{cases} 
\frac{q_{i,j}}{\sum_{k \neq i} q_{i,k}} & \text{if } i \neq j \\
0 & \text{otherwise}
\end{cases}
\]
The rate matrix $Q_A$ of $(A_t; t \geq 0)$ is

$$q_{j,i} = \binom{j}{j-i+1} \int_0^1 x^{j-i-1}(1-x)^{i-1} \Lambda(dx)$$

$$q_{j,j} = -\sum_{i=1}^{j-1} q_{j,i}, \quad 2 \leq j \leq n$$

$$q_{j,i} = 0, \quad \text{otherwise}$$
Using eigenvectors and eigenvalues of $Q_A$

Eigenvalues of $Q_A$ are $\alpha(k) = q_{k,k}$

Left eigenvector $l^{(k)} = \left( l_1^{(k)}, \ldots, l_n^{(k)} \right)$

Right eigenvector $r^{(k)} = \left( r_1^{(k)}, \ldots, r_n^{(k)} \right)$

Obtained by recursions

$$l_j^{(k)} = \frac{q_{j+1,k} l_{j+1}^{(k)} + \cdots + q_{k,j} l_j^{(k)}}{q_{k,k} - q_{j,j}}, \quad 1 \leq j < k$$

$$r_j^{(k)} = \frac{q_{j,k} r_k^{(k)} + \cdots + q_{j,j-1} r_{j-1}^{(k)}}{q_{k,k} - q_{j,j}}, \quad 1 < k < j \leq n$$
The spectral decomposition of $Q_A$ yields the transition probabilities

$$G_{i,j}(t) \equiv \mathbb{P}[A_t = j | A_0 = i]$$

as

$$G_{i,j}(t) = \sum_{k=j}^{i} e^{-\alpha(k)t} r_{i}^{(k)} l_{j}^{(k)}$$
Transition probabilities $G_{i,j}$ for $i = 3$

$$G_{3,2}(t) = \frac{q_{3,2}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 \leq t, T_3 + T_2 > t]$$

$$G_{3,1}(t) = \frac{q_{3,2}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 + T_2 \leq t] + \frac{q_{3,3}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 \leq t]$$

$$G_{3,3}(t) = \mathbb{P}[T_3 > t]$$

and

$$G_{3,1}(t) + G_{3,2}(t) + G_{3,3}(t) = 1$$
An example with $\Lambda_\psi$

Process with infinitesimal parameters

$$q_{ij} = \binom{i}{j} \psi^{i-j+1}(1 - \psi)^{j-1}$$

For $i = 3$ we obtain, with $\alpha(k) \equiv \sum_{k=i-1}^{1} q_{ik}$

$$G_{3,2}(t) = \frac{3}{2} \left( e^{-\alpha(2)t} - e^{-\alpha(3)t} \right)$$

$$G_{3,1}(t) = 1 - \frac{3}{2} e^{-\alpha(2)t} + \frac{1}{2} e^{-\alpha(3)t}$$

$$G_{3,3}(t) = e^{-\alpha(3)t}$$
In general,

\[ G_{i,j}(t) = \sum_{c \in C_{i,j}} g_c(t), \quad 1 \leq j < i \]

in which \( c \) is a coalescence sequence; or a particular order of mergers in going from \( i \) to \( j \) sequences. Number of possible sequences is

\[ |C_{i,j}| = 2^{i-j-1} \]
\[ g_c(t) = \begin{cases} 
  p(c) \mathbb{P}[T(c) \leq t, T(c) + T_j > t] & \text{if } j > 1 \\
  p(c) \mathbb{P}[T(c) \leq t] & \text{if } j = 1 \\
  \mathbb{P}[T_i > t] & \text{if } j = i
\end{cases} \]

in which

\[
\mathbb{P}[T(c) \leq t, T(c) + T_j > t] = e^{-\alpha(j)t} \sum_{k=1}^{l} \frac{\gamma_k}{\beta(i_k, j)} \left(1 - e^{-\beta(i_k, j)t}\right)
\]

with \( \beta(i_k, j) \equiv \alpha(i_k) - \alpha(j) \);

and

\[
\mathbb{P}[T(c) \leq t] = \sum_{k=1}^{l} \gamma_k' \left(1 - e^{-\alpha(i_k)t}\right)
\]
Example: two species

The probability $\mathbb{P}[MC]$ of *monophyletic concordance* for two lines from each of two species, with $\alpha_X(k) = \sum_{1 \leq k \leq i-1} q_{ik}$ (for species $X$)

$$
\mathbb{P}[MC] = (1 - e^{-\alpha_A(2)t})(1 - e^{-\alpha_B(2)t}) \\
+ e^{-\alpha_A(2)t}(1 - e^{-\alpha_B(2)t}) \beta_{3,2}/3 \\
+ (1 - e^{-\alpha_A(2)t})e^{-\alpha_B(2)t} \beta_{3,2}/3 \\
+ e^{-\alpha_A(2)t} e^{-\alpha_B(2)t} \beta_{4,2} \beta_{3,2}/9
$$
Two species and two lines each

○ : $\Lambda_\psi$; \hspace{0.5cm} △ : $K + \Lambda_\psi$

○ : $Beta(2 - a, a)$
Two species and two lines each

○ : $\Lambda_{0.05}$;  △ : $K + \Lambda_{0.05}$;  ◇ : Beta$(0.95, 1.05)$;  + : $K$

(time $t$)
Two species and two lines each

○: $\Lambda_{0.99}$; △: $K + \Lambda_{0.99}$; ◆: $Beta(0.05, 1.95)$; +: $K$

\[ 0 \ 2 \ 4 \ 6 \ 8 \]
\[ 0.0 \ 0.2 \ 0.4 \ 0.6 \ 0.8 \ 1.0 \]

\[ \text{time } t \]
Two species and three lines each

○ : $\Lambda_\psi$;  △ : $K + \Lambda_\psi$

○ : $Beta(2 - a, a)$
Two species and three lines each

○ : $\Lambda_{0.05}$;  △ : $K + \Lambda_{0.05}$;  ◇ : $\text{Beta}(0.95, 1.05)$

time $t$
Two species and three lines each:

- ○: $\Lambda_{0.95}$
- △: $K + \Lambda_{0.95}$
- ◇: $Beta(0.05, 1.95)$
Recursive approach for $s$ species

Let $\tilde{n} = n_1 + \cdots + n_s$ in which $n_i$ denotes the number of ancestral lines for species $i$ in a population; and let $n = (n_1, \ldots, n_s)$

$$
\mathbb{P}[MC; n] = \sum_{k=2}^{\tilde{n}} \beta_{\tilde{n},k} \sum_{r=1}^{s} \mathbb{P}[MC; m] \binom{n_r}{k} / \binom{\tilde{n}}{k}
$$

in which $m = (n_1, n_2, \ldots, n_{r-1}, n_r - k + 1, n_{r+1}, \ldots, n_s)$ and

$$
\mathbb{P}[MC; (0, 0, \ldots, 0, 1)] = \mathbb{P}[MC; (0, 0, \ldots, 0, 1, 1)] = 1
$$
Three species and two lines each \((t_1 = 1, t_2 = 2)\)

- \(\circ: \Lambda_{0.05}\)
- \(\triangle: K + \Lambda_{0.05}\)
- \(\diamond: Beta(0.95, 1.05)\)

\[\psi = a - 1\]
Three species and two lines each (+ : $K$)

- $\psi = 0.05; \ a = 1.95; \ t_2 = 0.05 + t_1$
- $\psi = 0.95; \ a = 1.001; \ t_2 = 0.05 + t_1$
- $\psi = 0.95; \ a = 1.95; \ t_2 = 1 + t_1$
- $\psi = 0.95; \ a = 1.001; \ t_2 = 1 + t_1$

- $\circ : \Lambda_{0.05}$
- $\triangle : K + \Lambda_{0.05}$
- $\diamond : Beta(0.95, 1.05)$
Conclusions

- Probability of monophyletic concordance depends on parameters of multiple merger coalescent processes
- Presence of multiple mergers complicates computations
- Scaling time appropriately is important
Acknowledgments

EPSRC and Marsden Fund for funding