Kingman’s Unlabeled $n$-Coalescent

Raazesh Sainudiin

Biomathematics Research Centre

Department of Mathematics & Statistics, University of Canterbury, NZ,
‡ Department of Statistics, University of Oxford, UK, and
⊛ Biological Sciences, University of California, Irvine, USA.

Joint with: Peter Donnelly†, Bob Griffiths†, Gil McVean†, and Kevin Thornton⊛
Outline – Talk Outline

- The Coalescent Models
- Computationally Intensive Likelihoods
- A Paritllay-ordered Coalescent Experiments Graph
- Unlabeled $n$-Coalescent
- Results
- Summary
- Acknowledgments
Data and Model 1: $\phi \equiv \theta \in \Phi$, $\theta = 4N_e \mu$ (scaled mutation rate)

The Wright-Fisher Model – Random Mating, Constant Size, No Recombination/Selection

A Population of $N = 10$ homologous DNA seqns. of length $m$ and the Population History of site $i$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Data and Model 1:** $\phi \equiv \theta \in \Phi$, $\theta = 4N_e\mu$ (scaled mutation rate)

The Wright-Fisher Model – Random Mating, Constant Size, No Recombination/Selection

Ex: Data of 3 homologous DNA sequences at site $i$, its Population History and the Sample History of sampled individuals 1, 2, and 3.

| :1 2 3 |
| i: T T A |

MRCA of the population

MRCA of the sample

time
**Model 1:** $\phi \equiv \theta \in \Phi, \theta = 4N_e \mu$ (scaled mutation rate)

The Coalescent Approximation of the Wright-Fisher (W-F) Model (Kingman, 1982)

A Sample Coalescent Sequence or $c$-sequence ( \( \{\{1\}, \{2\}, \{3\}\}, \{\{1, 2\}, \{3\}\}, \{\{1, 2, 3\}\} \) )

and coalescent times or epoch times $t_i, i \in \{3, 2\}$.

- Offspring “choose” parents uniformly and independently in W-F model
- $\Pr(2 \text{ lineages coalesce in 1 generation}) = 1/N$
- $\Pr(2 \text{ lins. are distinct } > g \text{ gens.}) = (1 - 1/N)^g$
- Rescaled time $t$ is $g$ in units of $N$ gens. Then, $\Pr(2 \text{ lins. remain distinct } > t)$ is
  $$\left(1 - 1/N\right)^{\lfloor Nt\rfloor} \xrightarrow{N \to \infty} e^{-t}$$
- **Lineage Death Process:** In general, the R.V. $T_i$ that any pair of $i$ lineages coalesce is approximately exponentially distributed for large $N$.
  $$T_i \sim \text{Exponential} \left(\binom{i}{2}\right)$$
- **Uniform Binary Fusion** of two extant lineages.
Model 1: $\phi \equiv \theta \in \Phi$, $\theta = 4N_e\mu$ (scaled mutation rate)

The Coalescent Approximation of the Wright-Fisher (W-F) Model (Kingman, 1982)

The $n$-Coalescent is a continuous time Markov Chain on $C_n \equiv \cup_{i=1}^{n} C^i_n$, the set partitions of $\{1, \ldots, n\}$, with rates $q(c_h|c_g)$, $c_g, c_h \in C_n$:

$$q(c_h|c_g) = \begin{cases} 
-i(i-1)/2 & : \text{if } c_g = c_h \in C^i_n \\
1 & : \text{if } c_h \succ_c c_g \\
0 & : \text{o.w.}
\end{cases}$$

$c_h \prec_c c_g \iff c_h = c_g \setminus c_{g,j} \setminus c_{g,k} \cup (c_{g,j} \cup c_{g,k})$

A realization $c = (c_n, c_{n-1}, \ldots, c_1) \in C_n$

Superimpose indep. mutations

$\sim Poisson(\theta/2 \equiv 2N\mu)$

$\infty$-many-sites mutation model
Coalescent Space $\mathcal{A}_n \equiv C_n \otimes (0, \infty)^{n-1}$ when $n = 3$ (Model 1)
Realizations from $\mathcal{A}_n \equiv C_n \otimes (0, \infty)^{n-1}$ under Model 1, $n = 6, 32$
Model 2: $\phi \equiv (\theta, \nu) \in \Phi$, $\theta = 4N_e \mu$ (scaled mutn. rate), $\nu$ (exp. growth rate)

Figures 1-6 of M. Nordburg, Coalescent Theory, 2000
(1) Every directed acyclic subgraph of the POEG indexes a Martingale

(2) Each node of the POEG is a tri-sequential asymptotic family of Experiments
Likelihood

Likelihood, $P(D|\phi)$, is computed by Integrating Missing-Data:

$$\sum_{c \in C_n} \int_{t \in (0,\infty)^{n-1}} P(D|c, t, \phi) P(c, t|\phi) dt \ dc$$

Cardinalities of the state spaces of the standard $n$-coalescent on $C_n$ and the unlabeled $n$-coalescent on $F_n$ (to be seen in the sequel).

<table>
<thead>
<tr>
<th>$n$</th>
<th>4</th>
<th>10</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>C_n</td>
<td>$</td>
<td>15</td>
<td>$1.2 \times 10^5$</td>
<td>$8.5 \times 10^{23}$</td>
</tr>
<tr>
<td>$</td>
<td>F_n</td>
<td>$</td>
<td>5</td>
<td>42</td>
<td>$5.6 \times 10^3$</td>
</tr>
<tr>
<td>$</td>
<td>F_n</td>
<td>/</td>
<td>C_n</td>
<td>$</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Likelihood is computationally prohibitive at MSA/BIM Resolns.

**Exact Methods:**
- **MSA**
  - 10,000 Auto-validating i.i.d. Posterior Samples in MRS SY2006 – novel
  - (3/4 leaved phylogenetic tree spaces)
  - \( \approx 200 \text{ CPU sec for } n \leq 3 \),
  - \( \vdash ( \rightarrow \text{ impractical for } n > 4 \)

- **BIM**
  - Complete Recursion in PTREE G1980
  - (1 Locus, \( \theta = 10 \), C-Model 1)
  - \( \vdash ( \rightarrow \text{ out of stack for } n > 4 \)

**Approximate Methods:**
- **MSA**
  - MCMC in COALESCE KYF1998 : \( n < 200 \) & heuristic
- **BIM**
  - SIS in GENETREE GT1994 : \( L(\theta|v) \approx 4 \text{ CPU hrs} / \theta \)

**The Bottom Line:**
- Exact Genome Scanning at fine DNA resolution is currently impractical for \( n > 4 \)
- **A Solution:** Inference at coarser empirical resolutions, eg. SFS and its sub-experiments – novel
Let $v \in \mathcal{V}_n^m \equiv \{0, 1\}^{n \times m}$ be a BIM, then the SFS $x \equiv (x_1, \ldots, x_{n-1}) \in \mathcal{X}_n^m \equiv \{x \in \mathbb{Z}_+^{n-1} : \sum_{i=1}^{n-1} x_i \leq m\}$

$$x_i = N_i(v^T \cdot (1, 1, \ldots, 1)), \quad N_i(y_1, y_2, \ldots, y_s) = \sum_{j=1}^{s} 1_{\{i\}}(y_j), \quad i = 1, \ldots, n-1.$$
Coalescent Tree Shape, $f$-Sequence and Site Frequency Spectrum

\[ \tilde{c}^\wedge = ((\cdot,3 \cdot)_1 (\cdot,2 \cdot)) \]

\[ f^\wedge = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix} \]

\[ P(f^\wedge) = 1/3 \]

\[ \mathcal{F}_4 = \{ f^\wedge, f^\wedge \} \]
Examples of $c$-sequence $\rightarrow f$-sequence, when $n = 4$

**Transition-Diagram**

**Hasse-Diagram**

Ex 1:

$[\{1\}, \{2\}, \{3\}, \{4\}], [\{1, 2\}, \{3\}, \{4\}], [\{1, 2, 3\}, \{4\}], [\{1, 2, 3, 4\}] \rightarrow [\{4, 0, 0, 0\}, \{2, 1, 0, 0\}, \{1, 0, 1, 0\}, \{0, 0, 0, 1\}]$

Ex 2:

$[\{1\}, \{2\}, \{3\}, \{4\}], [\{1, 2\}, \{3\}, \{4\}], [\{1, 2, 3\}, \{4\}], [\{1, 2, 3, 4\}] \rightarrow [\{4, 0, 0, 0\}, \{2, 1, 0, 0\}, \{0, 2, 0, 0\}, \{0, 0, 0, 1\}]$
Kingman’s Unlabeled $n$-Coalescent

Consider, the integer partitions of $n$ with $i$ blocks:

$$\mathcal{F}_n^i \equiv \{ f_i \equiv (f_{i,1}, f_{i,2}, \ldots, f_{i,n}) \in \mathbb{Z}_+^n : \sum_{j=1}^{n} j f_{i,j} = n, \sum_{j=1}^{n} f_{i,j} = i \}.$$ 

where $f_{i,j}$ denotes the number of lineages subtending $j$ leaves at the $i$-th epoch.

**Proposition** (Kingman’s Unlabeled $n$-coalescent). *It is the continuous time Markov chain on $\mathcal{F}_n \equiv \bigcup_{i=1}^{n} \mathcal{F}_n^i$, the set of integer partitions of $n$, whose infinitesimal generator $q(f_h | f_g)$ for any two states $f_g, f_h \in \mathcal{F}_n$ is:*

$$q(f_h | f_g) = \begin{cases} 0 & : \text{if } f_h = f_g \text{ and } f_g \in \mathcal{F}_n^i \\ f_{g,j} f_{g,k} & : \text{if } f_h = f_g - e_j - e_k + e_{j+k}, j \neq k, f_g \in \mathcal{F}_n^i, f_h \in \mathcal{F}_n^{i-1} \\ (f_{g,j})(f_{g,j} - 1)/2 & : \text{if } f_h = f_g - e_j - e_k + e_{j+k}, j = k, f_g \in \mathcal{F}_n^i, f_h \in \mathcal{F}_n^{i-1} \\ -i(i-1)/2 & : \text{if } f_g = f_h, f_g \in \mathcal{F}_n^i \\ 0 & : \text{otherwise} \end{cases}$$

*Initial state: $f_n = (n, 0, 0, \ldots, 0)$ and absorbing state: $f_1 = (0, 0, \ldots, 1)$. Any realization of the chain is an $f$-sequence: $f = (f_n, f_{n-1}, \ldots, f_1) \in \mathcal{F}_n.$*
Kingman’s Unlabeled $n$-Coalescent

**Proposition** (Probability of an $f_i$). *The probability of an $f_i \in \mathbb{F}_n^i$ is:*

$$P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \left(\frac{1}{n-1}\right)^{i-1}$$
Kingman’s Unlabeled \( n \)-Coalescent

**Proposition** (Probability of an \( f_i \)). The probability of an \( f_i \in F_n \) is:

\[
P(f_i) = \frac{i!}{\prod_{j=1}^{i} f_{i,j}!} \left( \frac{n-1}{i-1} \right)^{-1}
\]

**Proposition** (Probability of an \( f \)-sequence).

\[
P(f) = \prod_{i=2}^{n} P(f_i|f_{i-1}) = \frac{2^{-\gamma(f)}}{(n-1)!} \prod_{i=2}^{n} \tilde{f}_i
\]

where,

- \( \gamma(f) \) is the number of distinctly-sized lineage splits
- \( \tilde{f}_i \) is the number of lineages at the beginning of the \( i \)-th epoch that subtend the same number of leaves as the lineage that was split then.
$c$-sequence, $c \in \mathcal{C}_n \rightarrow c$-shape, $\tilde{c} \in \tilde{\mathcal{C}}_n \rightarrow f$-sequence, $f \in \mathcal{F}_n$

\[
\tilde{c}^\wedge = ((\cdot,3 \cdot),1 (\cdot,2 \cdot))
\]

\[
f^\wedge = \begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 2 & 0 & 0 \\
2 & 1 & 0 & 0 \\
4 & 0 & 0 & 0
\end{pmatrix}
\]

$n = 4$

\[
\tilde{c}^\wedge = (((\cdot,3 \cdot),2 \cdot),1 \cdot)
\]

\[
f^\wedge = \begin{pmatrix}
0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 \\
2 & 1 & 0 & 0 \\
4 & 0 & 0 & 0
\end{pmatrix}
\]
$c$-sequence, $c \in \mathcal{C}_n \rightarrow c$-shape, $\tilde{c} \in \tilde{\mathcal{C}}_n \rightarrow f$-sequence, $f \in \mathcal{F}_n$

$n = 5$

\[
\tilde{c}^{(a)} = (((\cdot, \cdot, \cdot), 3 \cdot, 2 \cdot), 1 \cdot) \\
\begin{pmatrix}
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 \\
2 & 0 & 1 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

\[
\tilde{c}^{(c)} = ((\cdot, 4 \cdot), 1 ((\cdot, 3 \cdot), 2 \cdot)) \\
\begin{pmatrix}
0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 \\
1 & 2 & 0 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

\[
\tilde{c}^{(e)} = (((\cdot, 4 \cdot), 3 \cdot), 1 (\cdot, 2 \cdot)) \\
\begin{pmatrix}
0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 \\
2 & 0 & 1 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
The number of \(c\)-sequences corresponding to the given \(f\) is

\[
| F^{-1}(f) | = 2^{1-n} n! (n - 1)! P(f) = n! 2^{\neg(f)+1-n} \prod_{i=2}^{n} f_i
\]

Let \(\mathcal{I}(\tilde{c})\) be the number of cherries of a \(c\)-shape \(\tilde{c} \in \tilde{C}\).

\[
| \tilde{C}^{-1}(\tilde{c}) | = 2^{1-n} n! (n - 1)! P(\tilde{c}) = n! 2^{-\mathcal{I}(\tilde{c})} \quad (Tajima, 1983)
\]

The number of \(c\)-shapes corresponding to the given \(f\) is

\[
| \tilde{C}(F^{-1}(f)) | = 2^{-\mathcal{I}(f)} \prod_{i=2}^{n} f_i, \\
\mathcal{D}(f) \equiv n - 1 - \neg(f) - \mathcal{I}(f), \text{ the number of balanced splits that are not cherries.}
\]
Hasse Diagram of the Poset making $\mathcal{F}_n$ ($n = 4, \ldots, 12$)

Raazesh Sainudiin, Department of Mathematics & Statistics, University of Canterbury

www.math.canterbury.ac.nz/~r.sainudiin – p. 18/27
Simulating $f$-sequences: for SFS, Shape Stats, ...

1: input:
   1. scaled mutation rate $\theta$
   2. sample size $n$

2: output: a SFS sample $x$ from the $n$-coalescent

3: generate an $f$-sequence under the unlabeled $n$-coalescent

4: draw $t \sim T = (T_2, T_3, \ldots, T_n)$, where $T_i$'s are
   independently distributed as Exponential $\left(\binom{i}{2}\right)$

5: $l \leftarrow t^T \cdot f$ and $l_\cdot = \sum_{i=1}^{n-1}$

6: draw $x$ from Poisson-Multinomial distribution
   $e^{-\theta l_\cdot} (\theta l_\cdot) \sum_{i=1}^{n-1} x_i \prod_{i=1}^{n-1} \bar{l}_i / \prod_{i=1}^{n-1} x_i!

7: return: $x$
Various tree shape statistics are further summaries of the $f$-sequence

$\tilde{s}$-sequence or Aldous shape statistic (Aldous, 2001)

$\tilde{S}(f_n, f_{n-1}, \ldots, f_1) = \tilde{s} \equiv (\tilde{s}_n, \tilde{s}_{n-1}, \ldots, \tilde{s}_2) : \mathcal{F}_n \to \tilde{S}_n$:

$\tilde{s}_i \equiv (\tilde{s}_{i,1}, \tilde{s}_{i,2}) \equiv (\max (\|f\|_i), \min (\|f\|_i)) 2^{-1} \{0\} (\max (\|f\|_i) - \min (\|f\|_i))$

$\|f\|_i \equiv \{ j | f_{i,j} - f_{i-1,j} | \in \mathbb{N} : j \in \{1, 2, \ldots, n\} \}$.

$Q_n \equiv \{ Q_I(\tilde{s}) = q_I \equiv \sum_{i=n}^{2} \tilde{s}_{i,1} 1_I(\tilde{s}_{i,1}) : \tilde{S}_n \to Q_{I_n}, I \in 2^{\{2,3,\ldots,n\} \setminus \emptyset} \}$

$Q_{\{2,3,\ldots,n\}}(\tilde{s}) = q_{\{2,3,\ldots,n\}} = \sum_{i=n}^{2} \tilde{s}_{i,1}$ is the Sackin’s index

$Q_{\{2\}}/2 = q_{\{2\}}/2$ is the number of cherries

$(n^2 - 3n + 2)^{-1} \sum_{i=n}^{2} (\tilde{s}_{i,1} - 2\tilde{s}_{i,d})$ is the Colless’ index

Note: There are $2^{n-1} - 3$ others in the family $Q_n$
Likelihood of a Site Frequency Spectrum

Proposition (Likelihood of SFS). Let \( a \in \mathcal{A}_n \) be a given coalescent tree, \( c \) be its \( c \)-sequence, \( f = F(c) \) be its \( f \)-sequence, \( t \equiv (t_2, t_3, t_n) \in (0, \infty)^{n-1} \) be its epoch times and let

\[
l \equiv (l_1, \ldots, l_{n-1}) = t^f = \left( \sum_{i=2}^{n} t_i f_{i,1}, \ldots, \sum_{i=2}^{n} t_i f_{i,n-1} \right), \quad l_\cdot \equiv \sum_{i=2}^{n} l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\cdot}
\]

be its lineage lengths subtending 1, 2, \ldots, \( n - 1 \) leaves, the total tree-size, and relative lineage lengths respectively.
Likelihood of a Site Frequency Spectrum

**Proposition (Likelihood of SFS).** Let $a \in \mathcal{A}_n$ be a given coalescent tree, $c$ be its $c$-sequence, $f = F(c)$ be its $f$-sequence, $t \equiv (t_2, t_3, t_n) \in (0, \infty)^{n-1}$ be its epoch times and let

$$l \equiv (l_1, \ldots, l_{n-1}) = t^T f = \left( \sum_{i=2}^{n} t_i f_i, 1, \ldots, \sum_{i=2}^{2} t_i f_i, n-1 \right), \quad l_\bullet \equiv \sum_{i=2}^{n} l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\bullet}$$

be its lineage lengths subtending $1, 2, \ldots, n-1$ leaves, the total tree-size, and relative lineage lengths respectively.

$$P(x|\phi, a) = P(x|\phi, l = t^T f) = e^{-\theta l_\bullet} (\theta l_\bullet^S) \prod_{i=1}^{n-1} \frac{l_i^{x_i}}{x_i!}$$
Likelihood of a Site Frequency Spectrum

**Proposition (Likelihood of SFS).** Let \( a \in \mathcal{A}_n \) be a given coalescent tree, \( c \) be its \( c \)-sequence, \( f = F(c) \) be its \( f \)-sequence, \( t \equiv (t_2, t_3, t_n) \in (0, \infty)^{n-1} \) be its epoch times and let

\[
l \equiv (l_1, \ldots, l_{n-1}) = t^T f = \left( \sum_{i=2}^n t_i f_{i,1}, \ldots, \sum_{i=2}^n t_i f_{i,n-1} \right), \quad l^* \equiv \sum_{i=2}^n l_i, \quad \bar{l}_i \equiv \frac{l_i}{l^*}
\]

be its lineage lengths subtending 1, 2, \ldots, \( n - 1 \) leaves, the total tree-size, and relative lineage lengths respectively.

\[
P(x|\phi, a) = P(x|\phi, l = t^T f) = e^{-\theta l^*} (\theta l^*)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!
\]

\[
P(x|\phi) = \frac{1}{\prod_{i=1}^{n-1} x_i!} \sum_{f \in F_n} \sum_{f \in \mathcal{F}_n(x^\circ)} P(f) \left( \int_{t \in (0, \infty)^{n-1}} e^{-\theta l^*} (\theta l^*)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \right) P(t|\phi)
\]

where,

\[
F_n(x^\circ) \equiv \bigcup_{\{h:x_h^\circ = 1\}} \{ f \in \mathcal{F}_n : \sum_{i=1}^n f_{i,h} = 0 \}
\]

\[
X^\circ(x) = x^\circ \equiv (x_1^\circ, \ldots, x_{n-1}^\circ) \equiv (1_N(x_1), \ldots, 1_N(x_{n-1})) \in \{0, 1\}^{n-1}
\]
An Importance Sampler over $F_n^c(x^\otimes)$

**Proposition** (A Proposal over $F_n^c(x^\otimes)$). For a given $x \in X_m^n$, consider the following discrete time Markov chain on the augmented state space $\mathbb{F}_n \times \{0, 1\}^{n-1} \ni (f_h, z_h)$:

$$P^*((f_h, z_h)|(f_g, z_g)) = \begin{cases} P(f_h|f_g)/\Sigma(f_g, z_g) & : \text{if } (f_h, z_h) \prec_f (f_g, z_g), \\ 0 & : \text{otherwise} \end{cases}$$

where,

$$\Sigma(f_g, z_g) = \sum_{(j,k) \in H(f_g, z_g)} P(f_g - e_{j+k} + e_j + e_k|f_g),$$

$$H(f_g, z_g) = \{(j,k) : f_{g,j+k} > 0, 1 \leq j \leq \max\{\min\{\hat{g}, j + k - 1\}, \left\lceil \frac{j + k}{2} \right\rceil \} \leq k \leq j + k - 1\},$$

$$\hat{g} = \max\{i : z_{g,i} = 1\},$$

$$(f_h, z_h) \prec_f (f_g, z_g) \iff f_h = f_g + e_j + e_k - e_{j+k}, z_h = z_g - 1_{\{1\}}(z_{g,j})e_j - 1_{\{1\}}(z_{g,k})e_k$$

where, the initial state is $(f_1, X^\otimes(x)) = ((0, 0, \ldots, 1), x^\otimes)$ and the final absorbing state is $(f_n, (0, 0, \ldots, 0)) = ((n, 0, \ldots, 0), (0, 0, \ldots, 0))$. 

Raazesh Sainudiin, Department of Mathematics & Statistics, University of Canterbury  www.math.canterbury.ac.nz/~r.sainudiin – p. 22/27
Maximum *Aposteriori* Estimates of $\theta$ and $\nu$ by $\sum$ over $f \in F_n^c(x^{\otimes})$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\hat{\nu}$</th>
<th>$\hat{\theta}$</th>
<th>$(\hat{\theta}, \hat{\nu})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sqrt{se}$</td>
<td>$bs$</td>
<td>$C_{99%}$</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>30</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>19</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>12</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>11</td>
<td>64</td>
</tr>
</tbody>
</table>
Topological Unfolding of SFS and Tajima’s D when $n = 4$
Simulated Vs. Gen. Fisher’s Exact Test with Tajima’s D

P-values for Simulated Vs. Exact Tajima’s D Test (theta = 1, 10, 50)

Left panel: Distribution of p-values from the simulated test (left) and the generalized Fisher’s exact test (right) for three values of $\theta = \{1, 10, 50\}$ per 1000 bp with $n = 30$.

Right panel: The almost zero correlation of p-values between the two tests.
Limits on Inference from Finest Empirical Resolutions

Inference from Coarser Site Frequency Spectrum is Possible via a Collapsed Kingman’s $n$-coalescent Markov chain

Algebraic Geometry is useful to infer from classical summaries of SFS.

MSEs are smaller – the exponential growth model

Helps speed-up intensive SIS methods (Particle filtering on Experiment Graph)

Topological unfolding of SFS and D $\Rightarrow$ Tree-less Genome Scans are essentially meaningless

A Decision-theoretic formalism – partially-ordered coalescent experiments graph

Possible to generalize

Saves electricity and slows down global warming!
NSF/NIGMS grant DMS-02-01037 to Durrett, Aquadro, and Nielsen and
Research Fellow of the Royal Commission for the Exhibition of 1851.