Matthias Birkner
Weierstrass Institute for Applied Analysis and Stochastics

Inference for coalescents with multiple collisions

Ongoing project with J. Blath and M. Steinrücken, TU Berlin

WIAS Institutskolloquium, 17 December 2007

Leibniz Gemeinschaft

Outline

1. Some 'classical' mathematical population genetics
2. Coalescents with multiple collisions
3. Combinatorics of the infinitely-many sites mutation model
4. A Monte Carlo method for likelihood estimation
5. Illustration

Genetic variability at a 250bp piece of the mitochondrial cytochrome *b*-gene in a sample of 117 atlantic cod (a random subsample from the dataset described in E. Árnason, *Genetics* 2004)

	468	481	487	488	490	496	508	523	562	601	631	643	649	685	691
66	t	a	a	c	a	a	t	g	a	t	g	a	c	c	g
17	-	-	-	-	-	-	c	-	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-	a	-	-	-	-	-	t	-
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	t
1	-	-	-	-	-	-	-	-	-	-	-	-	-	t	-
2	-	-	-	t	-	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	a	-	-	-	g	-	t	-
1	-	-	-	-	-	-	-	-	-	-	-	-	t	-	-
1	-	-	-	-	g	-	c	-	-	-	-	-	-	-	-
1	-	-	-	-	-	g	-	-	-	-	-	-	-	-	-
1	-	-	g	-	-	-	-	-	-	-	a	-	-	-	t
1	-	-	-	-	-	-	c	-	g	-	-	-	-	-	-
1	g	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-	c	-	-	-	-	-
1	-	c	-	-	-	-	c	-	-	-	-	-	-	-	-

John Gillespie's 'Great obsession' of population genetics:

"What evolutionary forces could have lead to such divergence between individuals in the same species?"

John Gillespie's 'Great obsession' of population genetics:

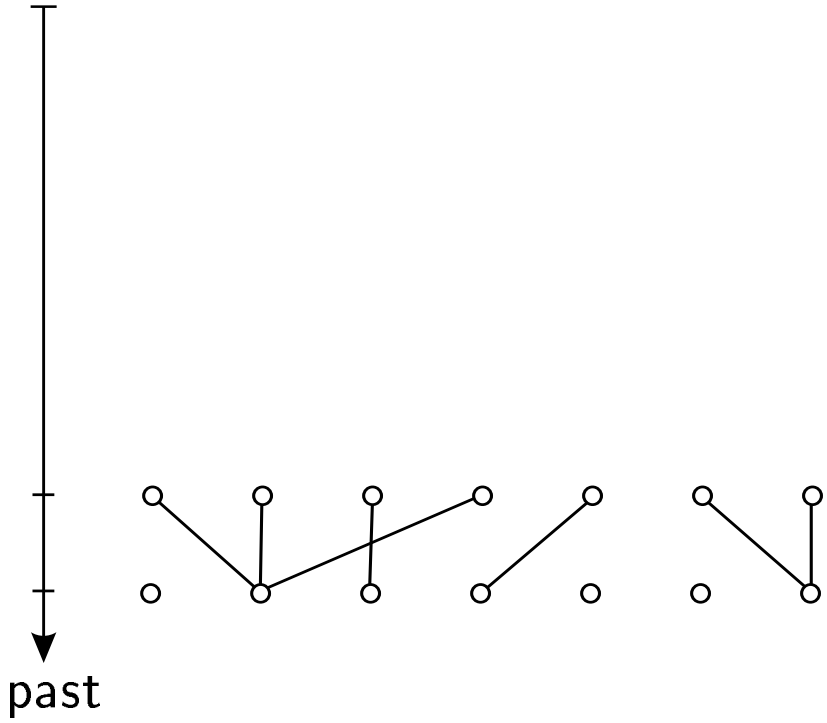
"What evolutionary forces could have lead to such divergence between individuals in the same species?"

A more humble obsession:

How can stochastic models help to understand genetic variability inside populations?

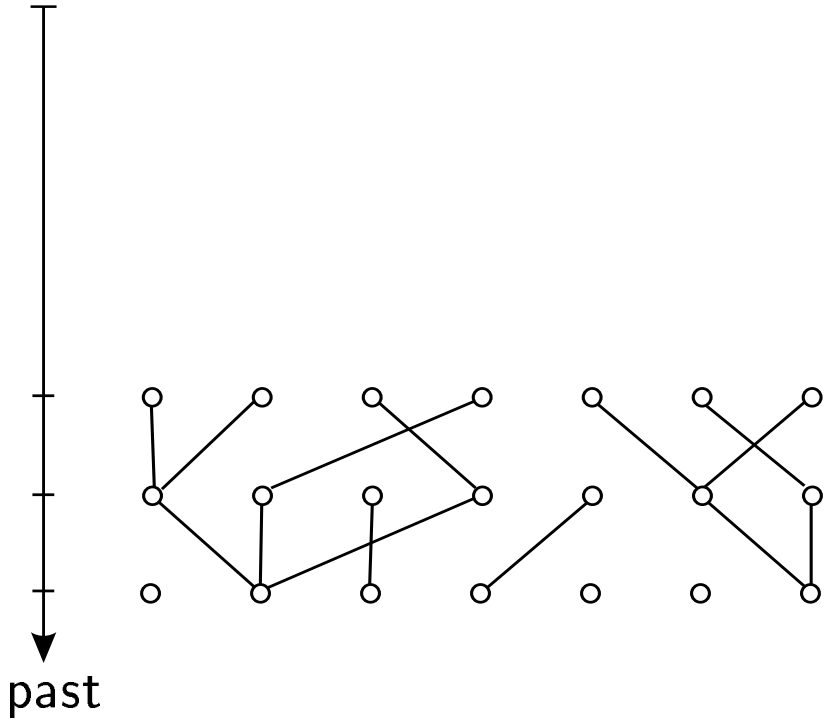
Wright-Fisher **model: A fundamental model for 'genetic drift'**

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



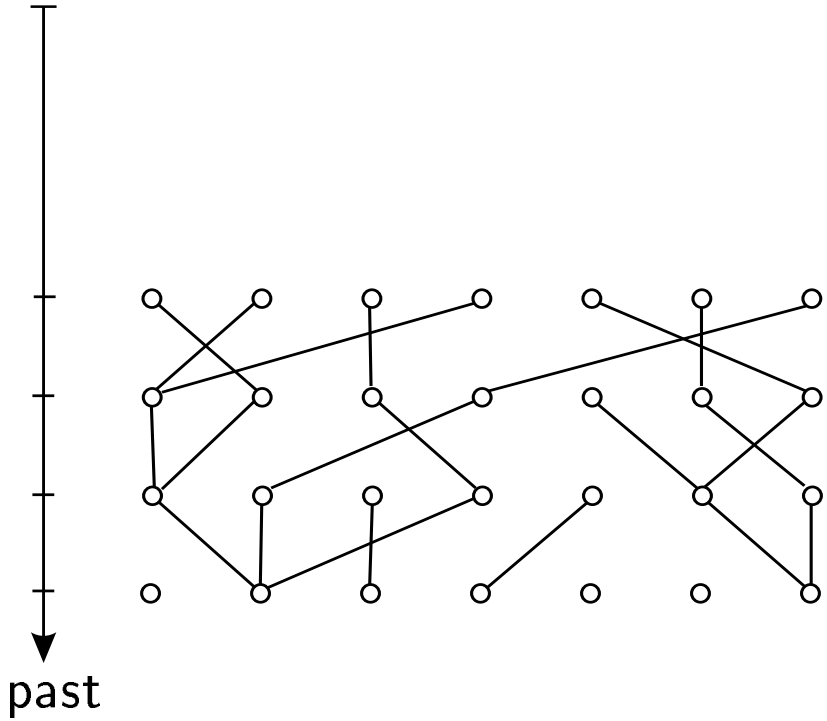
Wright-Fisher **model: A fundamental model for 'genetic drift'**

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



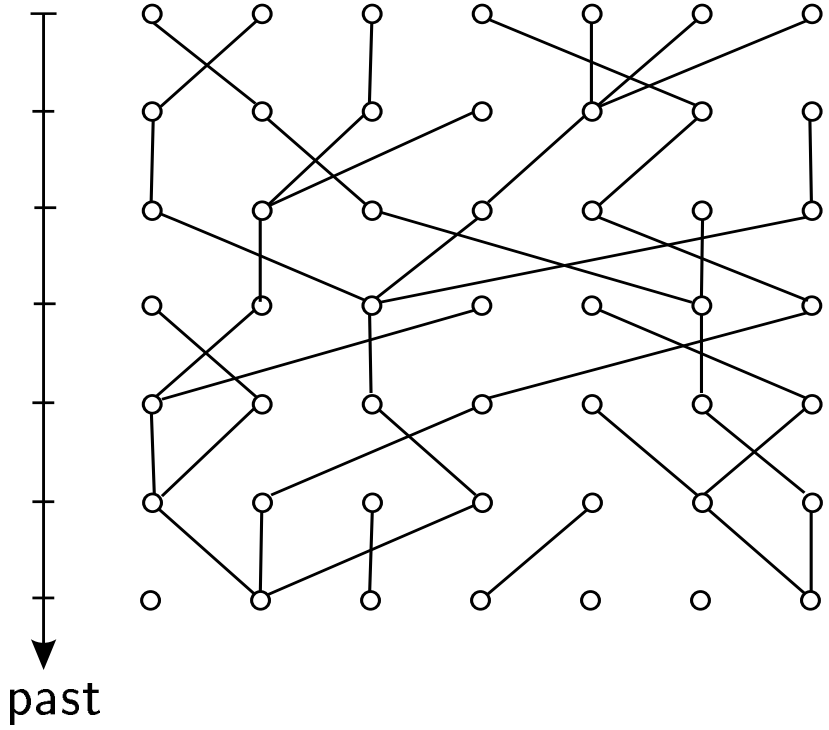
Wright-Fisher **model: A fundamental model for 'genetic drift'**

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



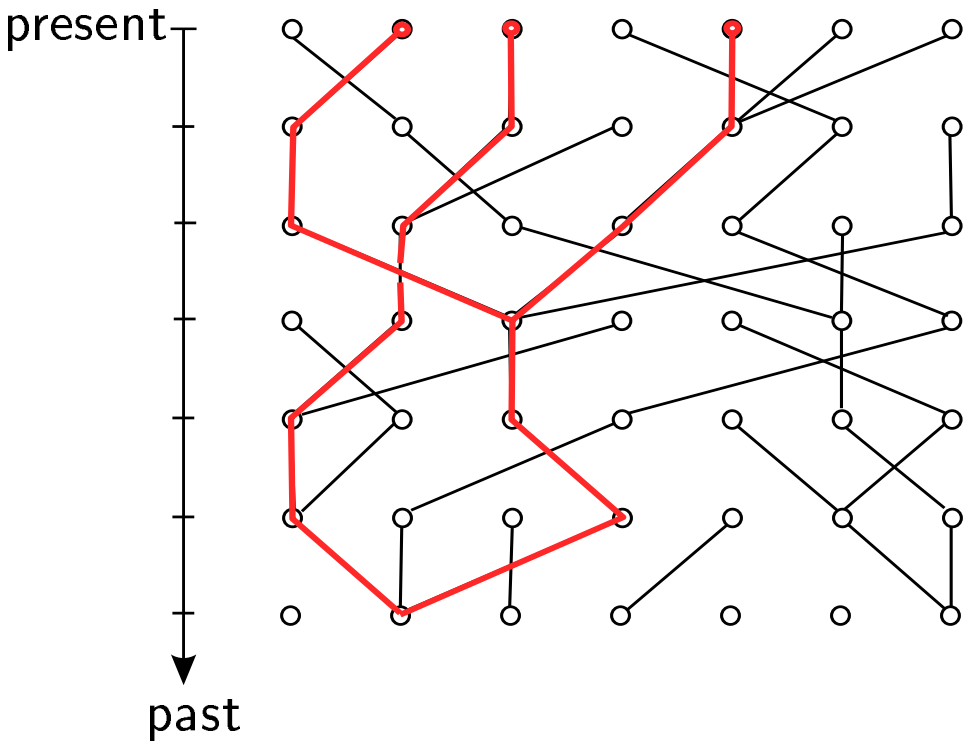
Wright-Fisher **model: A fundamental model for 'genetic drift'**

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



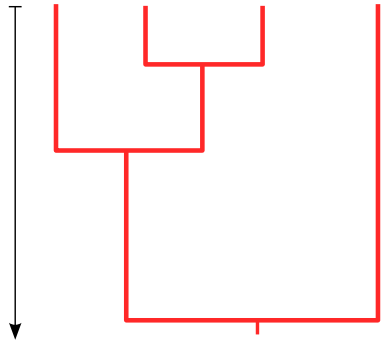
Genealogical point of view

Sample $n\,(\ll N)$ individuals from the 'present generation'

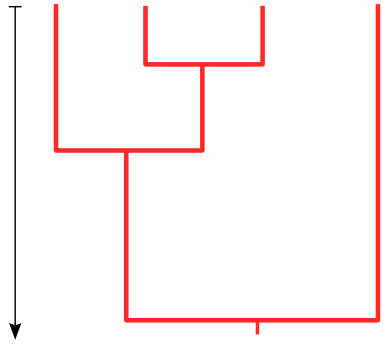


Kingman's coalescent

Theorem (Kingman, 1982)
In the limit $N \to \infty$, the genealogy of an n-sample, measured in units of N generations, is described by a continuous-time Markov chain where each pair of lineages merges at rate 1.



Kingman's coalescent



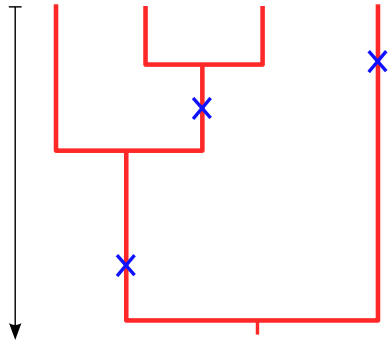
Theorem (Kingman, 1982)
In the limit $N \to \infty$, the genealogy of an n-sample, measured in units of N generations, is described by a continuous-time Markov chain where each pair of lineages merges at rate 1.

The same limit appears for any *exchangeable* offspring vectors

$$(\nu_1, \ldots, \nu_N), \qquad \text{(independent over generations)},$$

if time is measured in $\dfrac{N}{\sigma^2}$ generations, where $\sigma^2 = \lim\limits_{N\to\infty} \mathrm{Var}(\nu_1)$.

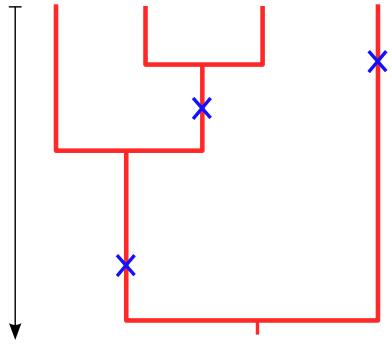
Kingman's coalescent: superimposing neutral types



Assume that the considered genetic types
do not affect their bearer's reproductive succes.

If as population size $N \to \infty$,
$\frac{N}{\sigma^2} \times$ mutation prob. per ind. per generation $\to r$,
the type configuration in the sample can be described by putting mutations with rate r along the genealogy.

Kingman's coalescent: superimposing neutral types



Assume that the considered genetic types
do not affect their bearer's reproductive succes.

If as population size $N \to \infty$,
$\frac{N}{\sigma^2} \times$ mutation prob. per ind. per generation $\to r$,
the type configuration in the sample can be described by putting mutations with rate r along the genealogy.

Kingman's coalescent is the *standard model* of mathematical population genetics.

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate
$\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$, where Λ is a finite measure on $[0,1]$.
(Sagitov, 1999; Pitman, 1999).

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate
$\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$, where Λ is a finite measure on $[0,1]$.
(Sagitov, 1999; Pitman, 1999).

Interpretation:
re-write $\lambda_{n,k} = \int_{[0,1]} x^{k}(1-x)^{n-k}\,\frac{1}{x^2}\Lambda(dx)$ to see:
at rate $\frac{1}{x^2}\Lambda([x, x+dx])$, an '$x$-resampling event' occurs.

Thinking forwards in time, this corresponds to an event in which the fraction x of the total population is replaced by the offspring of a single individual.

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate
$\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$, where Λ is a finite measure on $[0,1]$.
(Sagitov, 1999; Pitman, 1999).

Interpretation:
re-write $\lambda_{n,k} = \int_{[0,1]} x^{k}(1-x)^{n-k}\,\frac{1}{x^2}\Lambda(dx)$ to see:
at rate $\frac{1}{x^2}\Lambda([x, x+dx])$, an '$x$-resampling event' occurs.

Thinking forwards in time, this corresponds to an event in which the fraction x of the total population is replaced by the offspring of a single individual.

Note: $\Lambda = \delta_0$ corresponds to Kingman's coalescent.

Cannings' models in the 'domain of attraction of a Λ-coalescent'

Fixed population size N, *exchangeable* offspring numbers in one generation

$$\big(\nu_1, \nu_2, \ldots, \nu_N\big).$$

Sagitov (1999), Möhle & Sagitov (2001) clarify under which conditions the genealogies of a sequence of exchangeable finite population models are described by a Λ-coalescent:

- $c_N :=$ pair coalescence probability over one generation $\longrightarrow 0$
  $\big(\, c_N = \frac{1}{N-1}\mathbb{E}[\nu_1(\nu_1 - 1)] \,\big)$
- two double mergers asymptotically negligible compared to one triple merger
- $Nc_N\mathbb{P}\big(\text{a given family has size} \geq Nx\big) \sim \int_x^1 y^{-2}\Lambda(dy)$

Time is measured in $1/c_N$ generations (in general $\neq 1/\text{pop. size}$)

A 'heavy-tailed' Cannings model

Haploid population of size N. Individual i has X_i *potential offspring*,
$X_1, X_2, \ldots, X_N$ are i.i.d. with mean $m := \mathbb{E}\big[X_1\big] > 1$,
$\mathbb{P}\big(X_1 \geq k\big) \sim$ Const. $\times\, k^{-\alpha}$ with $\alpha \in (1, 2)$.
Note: infinite variance.

Sample N without replacement from all potential offspring to form the next generation.

A 'heavy-tailed' Cannings model

Haploid population of size N. Individual i has X_i *potential offspring*,
$X_1, X_2, \ldots, X_N$ are i.i.d. with mean $m := \mathbb{E}\big[X_1\big] > 1$,
$\mathbb{P}\big(X_1 \geq k\big) \sim$ Const. $\times\, k^{-\alpha}$ with $\alpha \in (1,2)$.
Note: infinite variance.

Sample N without replacement from all potential offspring to form the next generation.

Theorem (Schweinsberg, 2003)
Let $c_N =$ prob. of pair coalescence one generation back in N-th model.
$c_N \sim$ const. $N^{1-\alpha}$, measured in units of $1/c_N$ generations, the genealogy of a sample from the N-th model is approximately described by a Λ-coalescent with $\Lambda = \mathrm{Beta}(2-\alpha, \alpha)$.
$\Big(\mathrm{Beta}(2-\alpha,\alpha)(dx) = \mathbf{1}_{[0,1]}(x) \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}\, dx \Big)$

Why $\Lambda = \mathrm{Beta}(2-\alpha, \alpha)$?

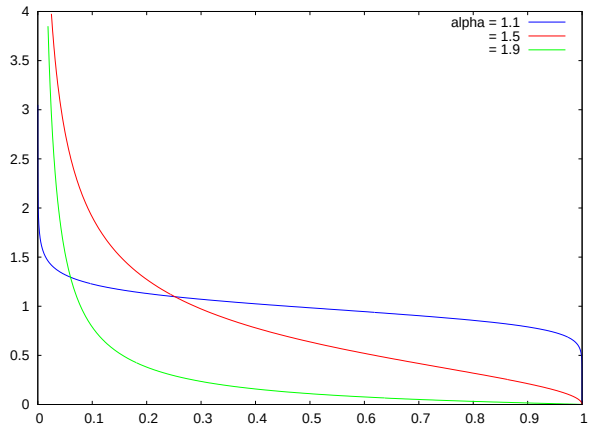
Heuristic argument:
Probability that first individual's offspring provides more than fraction y of the next generation, given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$)

Why $\Lambda = \mathrm{Beta}(2-\alpha, \alpha)$?

Heuristic argument:
Probability that first individual's offspring provides more than fraction y of the next generation, given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$)

$$
\begin{aligned}
\approx \quad & \mathbb{P}\Big(\frac{X_1}{X_1 + (N-1)m} \geq y \Big| X_1 \geq \varepsilon N\Big) \\
& = \mathbb{P}\Big(X_1 \geq (N-1)m\frac{y}{1-y} \Big| X_1 \geq \varepsilon N\Big) \\
& \sim \text{const.}\frac{(1-y)^\alpha}{y^\alpha} = \text{const.}' \,\mathrm{Beta}(2-\alpha, \alpha)([y,1]).
\end{aligned}
$$

The family $\mathrm{Beta}(2-\alpha,\alpha)$, $\alpha \in (1,2]$



- Kingman's coalescent is included as a boundary case: $\mathrm{Beta}(2-\alpha,\alpha) \to \delta_0$ weakly as $\alpha \to 2$.
- Smaller α means tendency towards more extreme resampling events.
- For $\alpha \leq 1$, corresponding coalescents *do not* come down from infinity.

- $\mathrm{Beta}(2-\alpha,\alpha)$-coalescents appear as genealogies of α-stable continuous mass branching process (via a time-change).

‘Meta-mathematic’ associations

Brownian motion	$\leftrightarrow$	Kingman’s coalescent
$\cap$		$\cap$
Stable processes	$\leftrightarrow$	Beta$(2-\alpha,\alpha)$-coalescents
$\cap$		$\cap$
General Lévy processes	$\leftrightarrow$	General Λ-coalescents

Asymptotics of the frequency spectrum

Consider an n-$\mathrm{Beta}(2-\alpha,\alpha)$-coalescent, mutations at rate r according to the *infinitely-many-sites* model (assuming known ancestral types). Let

$$\begin{aligned} M(n) &:= \#\text{total number of mutations in the sample,} \\ M_k(n) &:= \#\text{number of mutations affecting exaktly } k \text{ samples,} \end{aligned}$$

$k = 1, 2, \ldots, n-1$.

Theorem (Berestycki, Berestycki & Schweinsberg, 2005–)

$$\frac{M(n)}{n^{2-\alpha}} \to r\frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}, \quad \frac{M_k(n)}{n^{2-\alpha}} \to r\alpha(\alpha-1)^2\frac{\Gamma(k+\alpha-2)}{k!}$$

in probability as $n \to \infty$.

Asymptotics of the frequency spectrum

Consider an n-$\mathrm{Beta}(2-\alpha,\alpha)$-coalescent, mutations at rate r according to the *infinitely-many-sites* model (assuming known ancestral types). Let

$$\begin{aligned} M(n) &:= \#\text{total number of mutations in the sample,} \\ M_k(n) &:= \#\text{number of mutations affecting exaktly } k \text{ samples,} \end{aligned}$$

$k = 1, 2, \ldots, n-1$.

Theorem (Berestycki, Berestycki & Schweinsberg, 2005–)

$$\frac{M(n)}{n^{2-\alpha}} \to r\frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}, \quad \frac{M_k(n)}{n^{2-\alpha}} \to r\alpha(\alpha-1)^2\frac{\Gamma(k+\alpha-2)}{k!}$$

in probability as $n \to \infty$.

Thus $M_1(n)/M(n) \approx 2-\alpha$ for n large, which suggests

$$\widehat{\alpha}_{\mathrm{BBS}} := 2 - \frac{M_1(n)}{M(n)}$$

as an estimator for α.

A likelihood approach

If the observations had been generated by putting mutations at rate $r > 0$ on a realisation of a certain Λ-coalescent (from some class, e.g., Beta$(2-\alpha, \alpha)$), for which $(\hat{\Lambda}, \hat{r})$ is

$$\mathbb{P}_{\Lambda,r}(\text{observations}) \quad \text{maximal?}$$

Infinitely-many-sites model

An infinite sequence of completely linked sites, mutations always hit a new site

Example:

	segr. site			
Seq.	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	0	0	1	1
4	0	0	1	1
5	0	0	1	0

(0=wild type, 1=mutant

assume known ancestral types)

Obs. fit IMS $\Longleftrightarrow$ no sub-matrix $\begin{matrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{matrix}$ (nor row permutation).

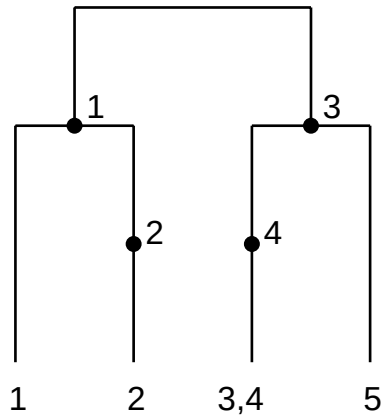
Infinitely-many-sites model, II

If the infinitely-many-sites model applies, the observations correspond to a unique rooted perfect phylogeny (or 'genetree').

Sequences,

	segr. site			
Seq.	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	0	0	1	1
4	0	0	1	1
5	0	0	1	0

Genetree,



obs. types

type	multiplicity
(1, 0)	1
(2, 1, 0)	1
(4, 3, 0)	2
(3, 0)	1

Construct e.g. using Gusfield's (1991) algorithm.

Note: purely combinatorial, does not depend on a probabilistic model for the observations.

"Naive approach"

We have

$$p_{\Lambda,r}(T,\mathbf{n}) = \sum_{\mathcal{T}\in\mathcal{C}_{T,\mathbf{n}}} \mathbb{P}_{\Lambda,r}\big(\text{marked geneal. tree of } n\text{-sample} = \mathcal{T}\big),$$

where $\mathcal{C}_{T,\mathbf{n}}$ are all marked coalescent trees compatible with the obsrvations.

Problem: Too many trees!

Recursions for tree probabilities (B. & Blath, 2007)

T a tree of (ordered) types, type multiplicity vector $\mathbf{n}$.

$$p_{\Lambda,r}(T,\mathbf{n}) = \frac{1}{r_n}\sum_{i:\,n_i\geq 2}\sum_{k=2}^{n_i}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1}p_{\Lambda,r}(T,\mathbf{n}-(k-1)\mathbf{e}_i)$$
$$+\frac{r}{r_n}\sum_{\substack{k:n_k=1,x_{k0}\,\text{distinct}\\ s(x_k)\neq x_j\,\forall j}} p_{\Lambda,r}(s_k(T),\mathbf{n})$$
$$+\frac{r}{r_n}\sum_{\substack{k:n_k=1,x_{k0}\\ \text{distinct}}}\sum_{j:s(x_k)=x_j}(n_j+1)p_{\Lambda,r}(r_k(T),r_k(\mathbf{n}+\mathbf{e}_j)).$$

where $\mathbf{e}_j$: j-the unit vector, $s_k(T)$: removes first coordinate of k-th sequence in $\mathbf{n}$, $r_k(T)$: removes k-th sequence from T, x_{k0} 'distinct': $\Longleftrightarrow$ $x_{k0}\neq x_{ij},\forall(x_1,\ldots,x_d)$ und $(i,j)\neq(k,0)$, $r_n = rn+\sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}$.

Extendes Ethier & Griffiths (1987) to Λ-case and Möhle (2005) to IMS.

Note: true recursion in *complexity* of $(T,\mathbf{n})$.

Markov chains and linear equations

$|S| < \infty$, (q_{xy}) transition kernel S, $f : S \to \mathbb{R}$.

$$u(x) = f(x) \sum_{y \in S} q_{xy} u(y), \quad x \in S' \subset S$$

with given boundary values on $S \setminus S'$.
X a q-Markov chain, $\tau := \min\{k : X_k \notin S'\}$.

If $\tau \leq K$ for a fixed $K < \infty$,

$$u(x) = \mathbb{E}_x \Big[\prod_{i=0}^{\tau} f(X_i) \Big].$$

A Monte-Carlo method

Unsing this and the recursion for $p_{\Lambda,r}$:

$$p_{\Lambda,r}(T,\mathbf{n}) = \mathbb{E}_{(T,\mathbf{n})}\Big[\prod_{i=0}^{\tau} f(X_i)\Big]$$

for a suitable Markov chain on type trees with multiplicities (analogous to Griffiths & Tavaré, 1994).

- Unbiased estimate $\widehat{p}_{\Lambda,r}(T,\mathbf{n})$ via independent runs
- Finite runtime: complexity of $(T,\mathbf{n})$ (:=#mutations + sample size) decreases in each step
- Can view chain as an integral on "(Λ-)coalescent histories"

The Monte-Carlo method (details to be glossed over)

Transition mechanism

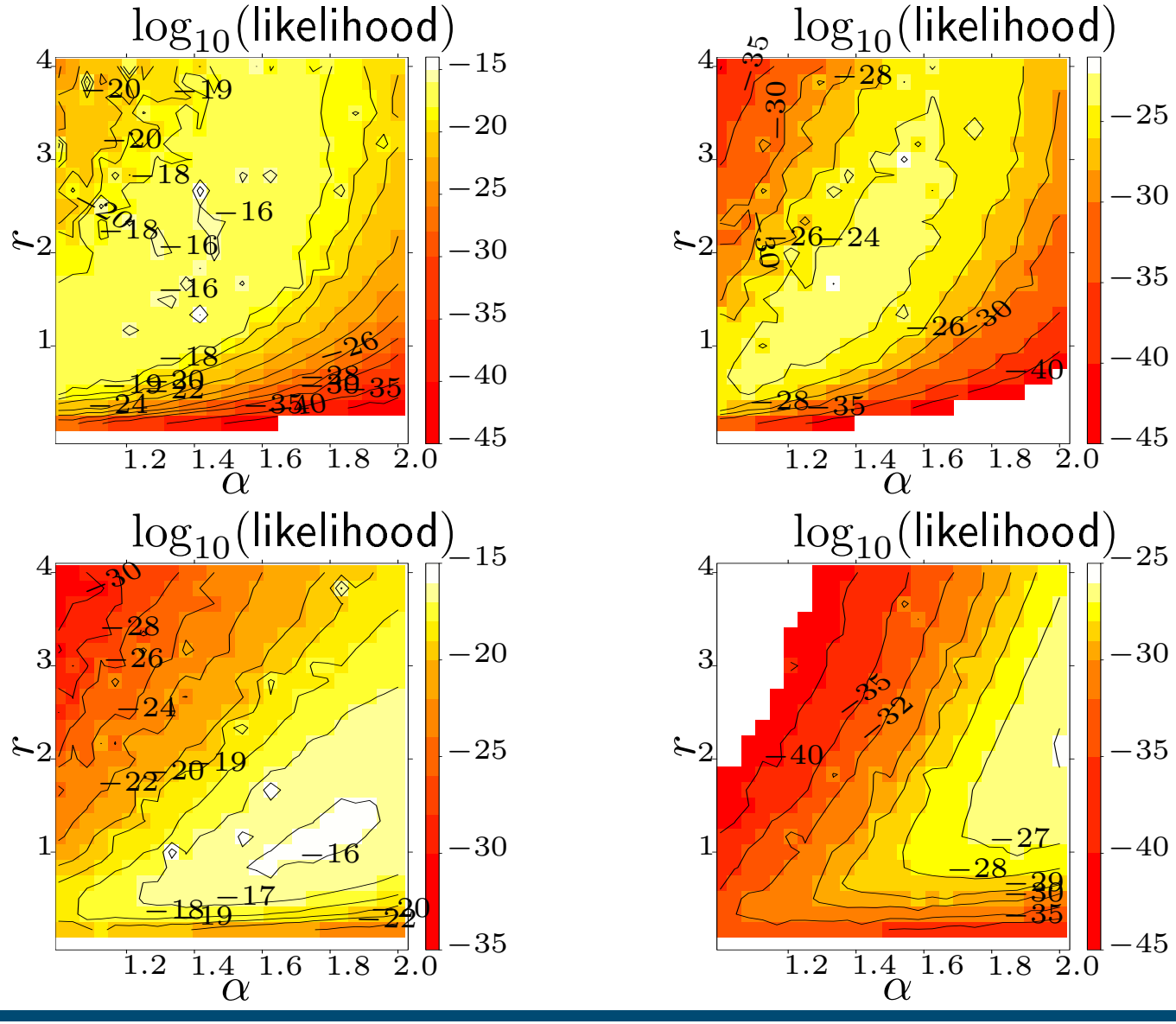
$$(T,\mathbf{n}) \to \begin{cases} (s_k(T),\mathbf{n}) & \text{w. p. } \frac{1}{r_n f(T,\mathbf{n})} r \text{ if} \\ & n_k = 1, x_{k0} \text{ distinct, } s(x_k) \neq x_j \forall j, \\ (r_k(T), r_k(\mathbf{n}+\mathbf{e}_j)) & \text{w. p. } \frac{1}{r_n f(T,\mathbf{n})} r(n_j+1) \text{ if} \\ & n_k = 1, x_{k0} \text{ distinct, } s(x_k) = x_j, \\ (T, \mathbf{n}-(k-1)\mathbf{e}_i) & \text{w. p. } \frac{1}{r_n f(T,\mathbf{n})} \binom{n}{k} \lambda_{n,k} \frac{n_i-k+1}{n-k+1} \text{ if } 2 \le k \le n_i, \end{cases}$$

where

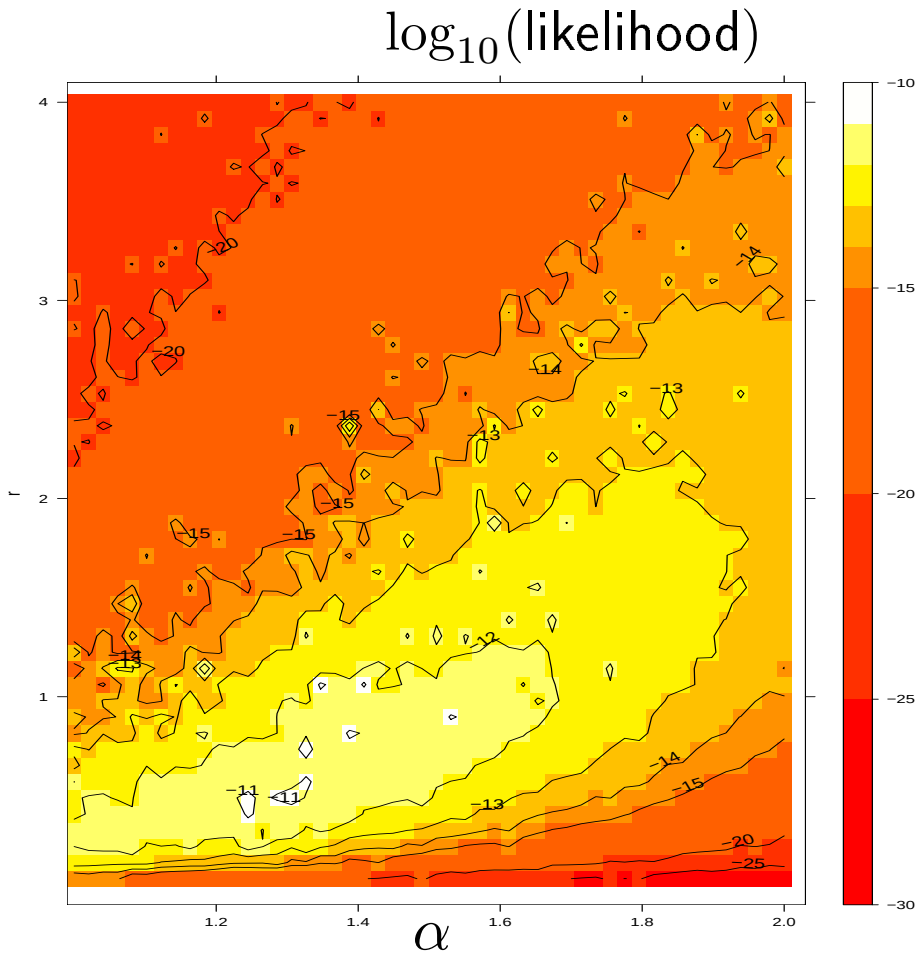
$$r_n f(T,\mathbf{n}) = \sum_{\substack{k: n_k=1, x_{k0} \text{distinct} \\ s_k(x_k) \neq x_j \forall j}} r + \sum_{\substack{k: n_k=1, x_{k0} \\ \text{distinct}}} \sum_{j: s_k(x_k)=x_j} r(n_j+1)$$

$$+ \sum_{1 \le i \le d:\, n_i \ge 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i-k+1}{n-k+1}.$$

Simulated datasets: $\alpha = 1.25,\ 1.5,\ 1.75,\ 2,\ r = 2.0$



$\hat{p}_{\mathrm{Beta}_{(2-\alpha,\alpha),r}}(\mathrm{data})$ for the cod sample



Maximum at $\hat{\alpha} = 1.3$, $\hat{r} = 0.7$. $\widehat{\alpha}_{\mathrm{BBS}} = 2 - 9/14 \approx 1.36$.

Further issues

- Reduce variance of estimator via importance sampling?
- Properties of estimators?
- Interplay of demographic stochasticity and recombination, "Λ-ancestral recombination graph"?
- More general mutation models, unknown ancestral types
- Selection, population substructure
- ...

`beta genetree` is available (under GNU General public licence) from

`http://www.wias-berlin.de/people/birkner/bgt/`