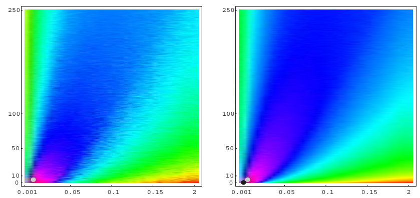
Exactly Approximate Bayesian Computations

Jim Booth[†], Raaz Sainudiin[‡], Mike Stillman[⋆] and Kevin Thornton[⊛]

[‡] Department of Statistics, University of Oxford,
Departments of [†]Biological Statistics and Computational Biology, [⋆]Mathematics and [⊛]Genetics, Cornell University

Trailer 1: Particle Filtering on Partially Ordered Experiments Graph – Donnelly & Griffiths

Trailer 2: A Markov Bases Family for Topologically-constrained SFS – Thornton & Yoshida



Outline

- The Dualistic Context of This Discourse:
  - **Tradition**: English Empiricism
  - **Universe of Hypotheses**: Popper's Falsifiability
  - **Internal Consistency** : Aristotelean Logic(s)
  - **Subject**: Mathematizable Statistical Genetics
  - **Engineering Constraints**: Resource-limited Info. Proc.
  - **Title**: Exactly Approximate Bayesian Computation
- Approximate Bayesian Computation
- A Coalescent Model and Associated Sample Spaces
- An Exactly Approximate Bayesian Computation – Rejection Methods I (Tavare's notes p. 3)
- Some Results, Summary, Extensions

Approximate Bayesian Computation – Motivation

Full likelihood methods are computationally prohibitive:

- Evaluation of the full likelihood function over the parameter space Ψ from $n = 90$ DNA sequences d_o is computationally intense – up to 4 hours per $\psi \in \Psi$ for the standard coalescent using Sequential Importance Sampling methods (Griffiths and Tavare, 1994)
- Computational time is prohibitive for complex models, for e.g. demographically structured coalescent

Approximate Bayesian Computation – Motivation

Full likelihood methods are computationally prohibitive:

- Evaluation of the full likelihood function over the parameter space Ψ from $n = 90$ DNA sequences d_o is computationally intense – up to 4 hours per $\psi \in \Psi$ for the standard coalescent using Sequential Importance Sampling methods (Griffiths and Tavare, 1994)
- Computational time is prohibitive for complex models, for e.g. demographically structured coalescent

A Practical Solution: Inference from Summaries:

- Let b'_o be a summary of the full data d_o
- Infer ψ from $P(\psi|b'_o) \simeq P(\psi|d_o)$
- b'_o NOT sufficient for $\psi \Longrightarrow$ ABC (Marjoram et al., 2003).

Approximate Bayesian Computation – A Simple Algorithm

(a) DRAW parameter ψ from the PRIOR $P(\psi)$

(b) SIMULATE an ancestral recombination graph (ARG) a with mutations m according to ψ and obtain data d

(c) SUMMARIZE d by b'

(d) ACCEPT ψ IF $||b', b'_o|| \leq \epsilon$, ELSE REJECT ψ

ITERATE (a)-(d) until you have enough accepted samples from an ϵ-specific approximation of $P(\psi|b'_o) \simeq P(\psi|d_o)$.

Approximate Bayesian Computation – A Simple Algorithm

(a) DRAW parameter ψ from the PRIOR $P(\psi)$

(b) SIMULATE an ancestral recombination graph (ARG) a with mutations m according to ψ and obtain data d

(c) SUMMARIZE d by b'

(d) ACCEPT ψ IF $||b', b'_o|| \leq \epsilon$, ELSE REJECT ψ

ITERATE (a)-(d) until you have enough accepted samples from an ϵ-specific approximation of $P(\psi|b'_o) \simeq P(\psi|d_o)$.

Variants on this basic theme include:

- Reweighting and smoothing through local regressions
- Bootstrap Filters, GLM, PCA, Projection Pursuits, ...
- Metropolis-Hastings, importance sampling, SMC, ...

Approximate Bayesian Computation – The ϵ Dilemma ! $\Rightarrow$ PCR !

The acceptance radius ϵ should be small, but not too small !

- Algorithm: Any ψ proposed from $P(\psi)$ is accepted if $||b', b'_o|| \leq \epsilon$. NOTE: When $\epsilon = 0$ we exactly get $P(\psi|b'_o)$
- ... BUT: $\epsilon \Rightarrow \downarrow$ acceptance rate and $\epsilon \Rightarrow P(\psi|b'_o) = P(\psi)$
- ... TUNE ϵ under the appropriate metric $||\cdot||$ to obtain the optimal trade-off between efficiency and accuracy

Approximate Bayesian Computation – The ϵ Dilemma ! $\Rightarrow$ PCR !

The acceptance radius ϵ should be small, but not too small !

- Algorithm: Any ψ proposed from $P(\psi)$ is accepted if $||b', b'_o|| \leq \epsilon$. NOTE: When $\epsilon = 0$ we exactly get $P(\psi|b'_o)$
- ... BUT: $\epsilon \Rightarrow \downarrow$ acceptance rate and $\epsilon \Rightarrow P(\psi|b'_o) = P(\psi)$
- ... TUNE ϵ under the appropriate metric $||\cdot||$ to obtain the optimal trade-off between efficiency and accuracy
- Question: Can we make ϵ to be exactly 0 ?
- Answer : YES! for several classical summaries

Model — $\psi \triangleq (\theta, \nu) \in \Psi$, $\theta = 4N_e\mu$ (scaled mutation rate), ν (exponential growth rate)

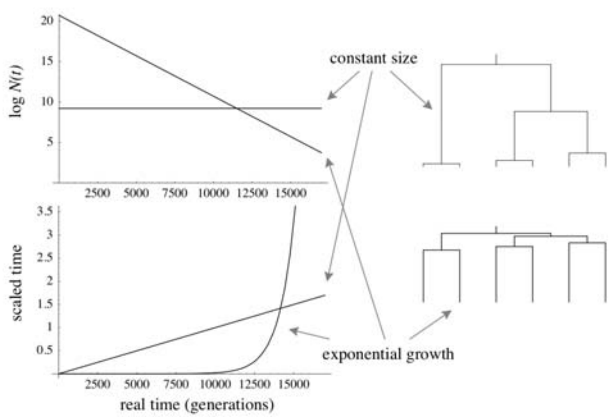
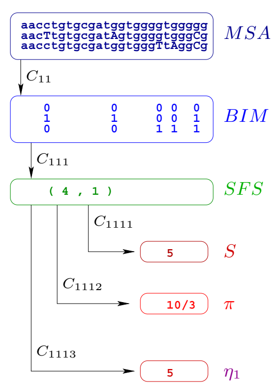
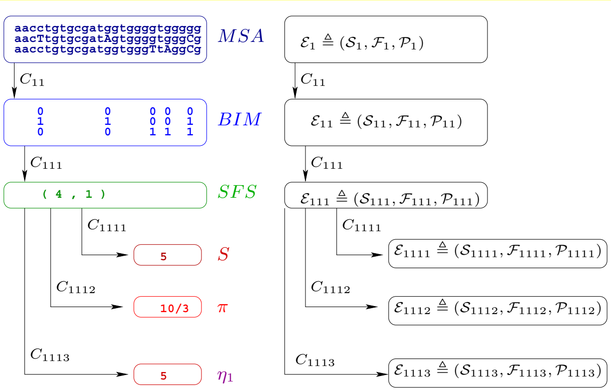


Figure 6. of M. Nordburg, Coalescent Theory, 2000

Coalescent Sample Spaces



Coalescent Sample Spaces – Partially Ordered Experiments Graph



$\mathcal{P}_\alpha \triangleq \{P_{\alpha\psi} : \psi \in \Psi\}$, where $\mathcal{F}_\alpha \triangleq \mathcal{F}_{\mathcal{S}_\alpha}$, $C_\beta : \mathcal{S}_\alpha \to \mathcal{S}_\beta$ and $\mathcal{E}_\alpha \geq \mathcal{E}_\beta \Leftrightarrow \exists\, C_\beta : \mathcal{S}_\alpha \to \mathcal{S}_\beta$.

Two Popular Linear Summaries of SFS $x \triangleq (x_1, \ldots, x_{n-1})$

Let $b = (S, \Pi)$ for fixed sample size n,

$$S \triangleq \sum_{i=1}^{n-1} x_i, \quad \pi \triangleq \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)x_i, \quad \Pi = \binom{n}{2}\pi.$$

Inference based on S and Π depends on the kernel of:

$$B \triangleq \begin{pmatrix} 1 & \ldots & 1 & \ldots & 1 \\ 1(n-1) & \ldots & i(n-i) & \ldots & n-1(n-(n-1)) \end{pmatrix}.$$

Consider the set of all SFS that exactly satisfy b.

It is the bounded non-empty polytope:

$$\Gamma_B^b \triangleq \{x \in \mathbb{Z}_+^{n-1} : Bx = b\}$$

Integrating over Γ_B^b

Question: Can we somehow sample from Γ_B^b ?
If we could, then we can do exactly ABC with $\epsilon = 0$.

Integrating over Γ_B^b

Question: Can we somehow sample from Γ_B^b ?
If we could, then we can do exactly ABC with $\epsilon = 0$.
Answer: YES! via computational commutative algebra.

Definition 0 (Markov Basis) *Let $\mathcal{M}$ be a finite subset of the kernel of $B \cap \mathbb{Z}^{n-1}$. Consider the undirected graph $\mathcal{G}_B^b$, such that (1) all nodes are lattice points in Γ_B^b and (2) edges between a node x and a node y are possible $\iff$ $x - y \in \mathcal{M}$. If the graph $\mathcal{G}_B^b$ is connected for all b, then $\mathcal{M}$ is called a Markov basis.*

Integrating over Γ_B^b

Question: Can we somehow sample from Γ_B^b ?
If we could, then we can do exactly ABC with $\epsilon = 0$.
Answer: YES! via computational commutative algebra.

Definition 0 (Markov Basis) *Let $\mathcal{M}$ be a finite subset of the kernel of $B \cap \mathbb{Z}^{n-1}$. Consider the undirected graph $\mathcal{G}_B^b$, such that (1) all nodes are lattice points in Γ_B^b and (2) edges between a node x and a node y are possible $\iff$ $x - y \in \mathcal{M}$. If the graph $\mathcal{G}_B^b$ is connected for all b, then $\mathcal{M}$ is called a Markov basis.*

Sampling Implication:
Monte Carlo Markov chains constructed with local moves from $\mathcal{M}$ are irreducible and can be made aperiodic, and are therefore ergodic on the finite state space Γ_B^b.

Some elements of a Markov Basis

A Markov basis for Γ_B^b with $n = 30$, computed using the software package for computational algebra `Macaulay 2` (Grayson and Stillman, 2004), had 520 elements.

Five of them are:

```
+0 +0 +0 +0 +0 +0 +0 +0 +0 -1 +1 +1 +0 +0 -1 +0 ... +0 +0
+2 -2 -2 +1 +0 +2 +0 -1 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
-3 +1 +4 -1 +0 +0 +0 +0 -1 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
+7 -9 +0 +0 +1 +0 +0 +1 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
+1 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 -1
```


Some elements of a Markov Basis – BUT, where are the ARGs ?

A Markov basis for Γ_B^b with $n = 30$, computed using the software package for computational algebra `Macaulay 2` (Grayson and Stillman, 2004), had 520 elements.

Five of them are:

```
+0 +0 +0 +0 +0 +0 +0 +0 +0 -1 +1 +1 +0 +0 -1 +0 ... +0 +0
+2 -2 -2 +1 +0 +2 +0 -1 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
-3 +1 +4 -1 +0 +0 +0 +0 -1 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
+7 -9 +0 +0 +1 +0 +0 +1 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 +0
+1 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 +0 ... +0 -1
```

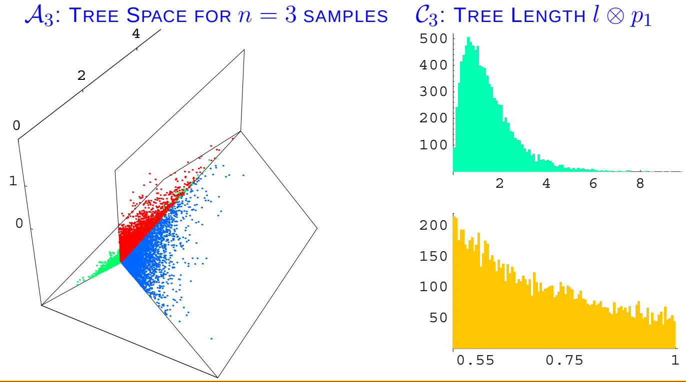
- OK, we can run MCMCs in $\Gamma_B^{b_o}$ if we initialize at x_o
- BUT, what is the target density over $\Gamma_B^{b_o}$? Where are the ARGs in this picture ?
- ARG-specific targets on $\Gamma_B^{b_o}$ are Poisson-Multinomials!

Sufficient Compression of $\mathcal{A}_n$ to $\mathcal{C}_n$

Let $a \in \mathcal{A}_n$ be an ARG and $\psi = (\theta, \nu)$. Let C map a into its total length l and relative lengths p_i that dictate mutations in SFS x :

$$C(a) = (l, p) : \mathcal{A}_n \to \mathcal{C}_n \triangleq \mathbb{R}_+ \otimes \triangle_{n-1}$$

$\mathcal{A}_3$: TREE SPACE FOR $n = 3$ SAMPLES $\quad$ $\mathcal{C}_3$: TREE LENGTH $l \otimes p_1$



The Exactly Approximate Posterior

$$P(b|\psi) = P(b,\psi)/P(\psi) \quad = \quad \int_{(l,p)\in\mathcal{C}_n} \sum_{x\in\Gamma_B^b} \mathfrak{PM}(x|\psi,l,p)P(l,p|\psi),$$

$$\text{where,} \quad \mathfrak{PM}(x|\psi,l,p) = e^{-\theta l}(\theta l)^S \prod_{i=1}^{n-1} p_i{}^{x_i} / \prod_{i=1}^{n-1} x_i!.$$

The Exactly Approximate Posterior

$$P(b|\psi) = P(b,\psi)/P(\psi) \quad = \quad \int_{(l,p)\in\mathcal{C}_n} \sum_{x\in\Gamma_B^b} \mathfrak{PM}(x|\psi,l,p)P(l,p|\psi),$$

$$\text{where,} \quad \mathfrak{PM}(x|\psi,l,p) = e^{-\theta l}(\theta l)^S \prod_{i=1}^{n-1} p_i{}^{x_i} / \prod_{i=1}^{n-1} x_i!.$$

Therefore, $P(\psi|b) \propto P(b|\psi)P(\psi)$

$$\approx \frac{1}{N}\sum_{j=1}^{N} \frac{1}{M} \sum_{h=1:x\in\Gamma_B^b}^{M} \mathfrak{PM}(x^{(h)}|\psi, l^{(j)}, p^{(j)}),\ (l^{(j)}, p^{(j)}) \sim P(l,p|\psi)P(\psi).$$

where, the sum over M $x^{(h)}$'s are obtained through a Metropolis-Hastings Markov chain (or an annealed SIS/popMCMC) on Γ_B^b with the ARG-specific target distribution $\mathfrak{PM}(x|\psi,l,p)$ and the Monte Carlo sum over N ARGs can be obtained from simulation under ψ.

Adding More Summaries of SFS x

- It is straightforward to add other popular linear summaries of the SFS x.
- For example, including $\eta_1 \triangleq x_1 + x_{n-1}$ to the previous two summaries S and Π yields the following matrix B':

$$B' \triangleq \begin{pmatrix} 1 & \dots & 1 & \dots & 1 \\ 1(n-1) & \dots & i(n-i) & \dots & n-1(n-(n-1)) \\ 1 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

- The cardinality of a Markov basis for $\Gamma_{B'}^b$ is 440 (smaller when compared to 520 for Γ_B^b conditioned by S and Π) when $n = 30$.

Estimating the scaled mutation rate θ

MSEs and Bias from 1000 Replicates simulated under $\theta = 10.0$

MSE of various estimators

n	$\widehat{\theta_W}$	$\widehat{\theta_\pi}$	ABC$_{S,\Pi}$	EABC$_{S,\Pi}$	SIS$_{BIM}$
10	23.19	31.86	26.30	19.13	12.57
30	12.88	25.81	14.83	10.54	6.94
90	7.90	24.98	7.45	6.33	4.07

Bias of various estimators

n	$\widehat{\theta_W}$	$\widehat{\theta_\pi}$	ABC$_{S,\Pi}$	EABC$_{S,\Pi}$	SIS$_{BIM}$
10	-0.10	-0.20	1.49	-0.58	-0.61
30	0.18	0.21	0.73	-0.13	-0.33
90	-0.12	0.011	0.21	-0.51	-0.55

- EABC$_{S,\Pi}$ is the Mean Tree Tajima-Waterson Estimator of θ given S and Π
- Just the first moment on $\mathcal{C}_n$, ie. mean tree length and mean relative time leading to singletons, doubletons, ..., '$(n-1)$tons for each ν

Estimating θ and growth rate ν

MSEs and Bias from 1000 Replicates simulated under $\theta = 10.0$, $\nu = 0.0$, $n = 30$

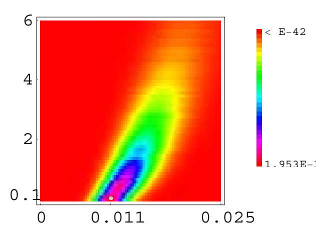
MSE (BIAS) of three estimators of θ and ν

parameter	ABC$_{S,\Pi}$	EABC$_{S,\Pi}$	EABC$_{S,\Pi,\eta_1}$
θ	$82.41(6.57)$	$50.18(4.14)$	$46.20(4.06)$
ν	$26.24(4.08)$	$11.75(2.13)$	$13.67(2.37)$

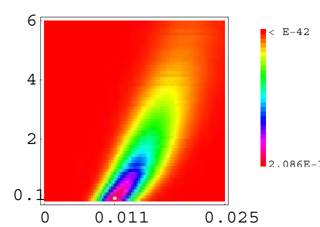
- ABC algorithm with smoothing and reweighting through local regressions was used with an acceptance radius $\epsilon = 0.001$.
- Computationally prohibitive to compare with SIS methods based on BIM
- Bottom line: Do exactly ABC when possible
- Rigorous 'zoning in' technique for intensive SIS methods

Approximate Posterior Density.

$P(\theta,\nu|S,\Pi)$



$P(\theta,\nu|S,\Pi,\eta_1)$

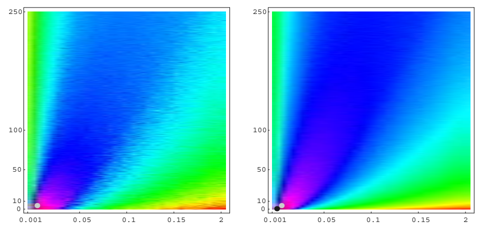


Shannon's information (Expected Negative Entropy $\triangleq E_P(\log(P))$) measure for $P(\theta,\nu|S,\Pi)$ and $P(\theta,\nu|S,\Pi,\eta_1)$ are -7.50989 and -7.49071, respectively. Thus, η_1 adds more information (0.0191824) by making $P(\theta,\nu|S,\Pi,\eta_1)$ more concentrated than $P(\theta,\nu|S,\Pi)$.

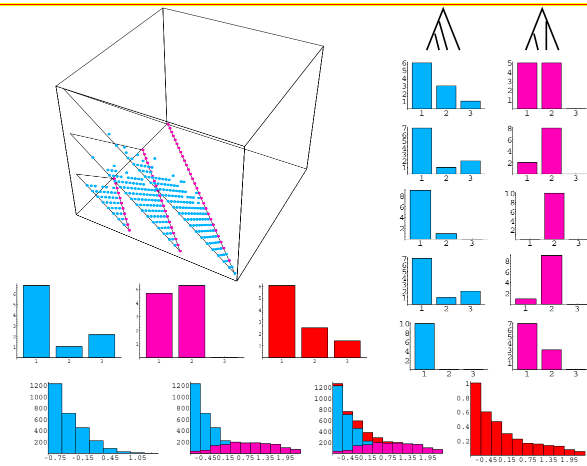
Independent M-H Sampling on $\mathcal{A}_n$ – A Poisson-Dirichlet Shave

$$\approx \frac{1}{N}\sum_{j=1}^{N}\frac{1}{M}\sum_{h=1:x\in\Gamma_B^b}^{M}\mathfrak{PM}(x^{(h)}|\psi, l^{(j)}, p^{(j)}),\ (l^{(j)}, p^{(j)}) \sim P(l,p|\psi)P(\psi).$$

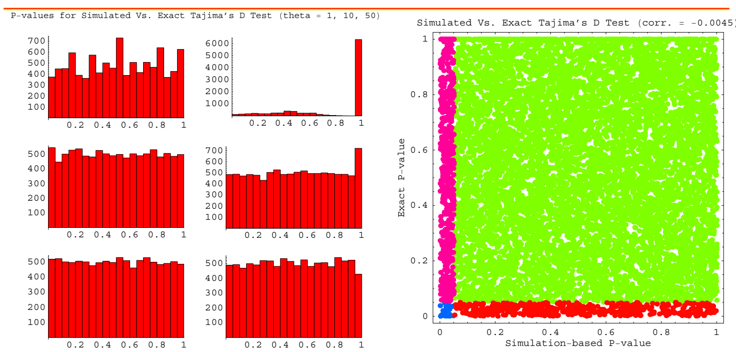
Can use N independent M-H samples of ARGs with independent proposal given by simulation under ψ and the target specified by the posterior distribution on $\mathcal{C}_n \triangleq \mathbb{R}_+ \otimes \triangle_{n-1}$ – a Poisson Dirichlet posterior based on observed S and x_o.



Topological Unfolding of SFS and Tajima's D when $n = 4$



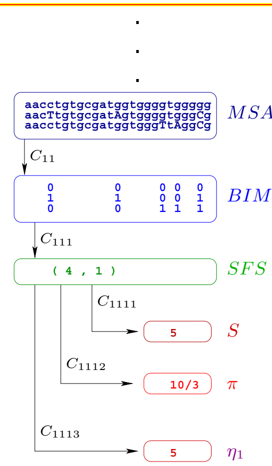
Simulated Vs. Gen. Fisher's Exact Test with Tajima's D



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.

Summary



- Full likelihood methods can be impractical
- Practical summary-based ABC methods: $\epsilon\ \epsilon\ \epsilon$
- Using a Markov basis for linear summaries of SFS we can do exactly ABC or ABC$_{\epsilon=0}$
- Can compute MB with standard algebraic software
- MSEs are smaller – the exponential growth model
- Helps 'zone in' prior to intensive SIS methods
- Decide between alternate sets of summaries through information measures – adding η_1 helps
- Can incorporate more information via Poisson-Dirichlet Shave
- Topological unfolding of SFS and D $\Rightarrow$ Tree-less Genome Scans
- A Decision-theoretic formalism – partially-ordered experiments graph

Discussion and Extensions

- Need not restrict to linear summaries of the SFS; structure (2D SFS), recombination (blocks of SFS $\otimes$ ARG Summaries)
- Hybrid Methods – add $\epsilon > 0$ ABC summaries
- Particle Filtering (SMC) along the filtration induced by the Partially-Ordered Experiments Graph (POEG)
- Disadvantage – for large $n > 200$ the Markov bases computations are exponentially slow (BUT only once!)
  - FIRST learn about optimal paths toward 'root' on POEG with smaller n – THEN do ABC with $\epsilon > 0$.
  - Information only grows logarithmically fast – $n > 200$ adds little information
- EG allows for (1) 'co-existence' of many methods, (2) analysis through LeCam's theory of experiments, (3) saves electricity and slows down global warming!

References

Grayson, D. R. and M. E. Stillman (2004). Macaulay 2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

Griffiths, R. C. and S. Tavare (1994). Ancestral inference in population genetics. *Statistical Science 9*, 307319.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA 100*, 15324–15328.

- **NSF/NIGMS grant DMS-02-01037 to Durrett, Aquadro, and Nielsen and**
- **Research Fellow of the Royal Commission for the Exhibition of 1851.**