Approximate Bayesian Computation: what, why and how

Simon Tavaré

DAMTP and Cambridge Research Institute

Cambridge Statistics Discussion Group, October 28 2008

Stochastic Computation in Biology

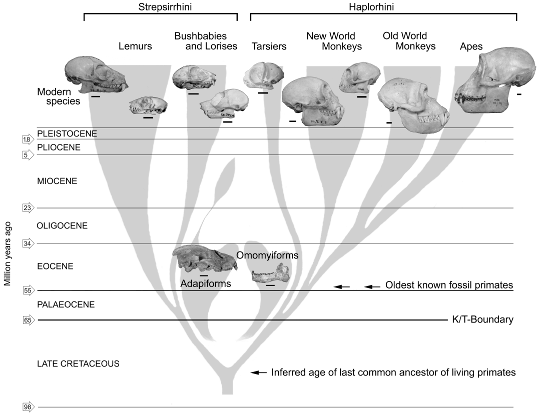
The Biological Motivation

Evolution on different time scales

- Tracking stem cells in an individual
- Ancestral history of humans
- Divergence time of primates from fossil record

Inference in the fossil record

Primate Evolution



Strepsirrhini

Haplorhini

Lemurs

Bushbabies and Lorises

Tarsiers

New World Monkeys

Old World Monkeys

Apes

Modern species

PLEISTOCENE

PLIOCENE

MIOCENE

OLIGOCENE

EOCENE

PALAEOCENE

LATE CRETACEOUS

Million years ago

1.8

5

23

34

55

65

98

Omomyiforms

Adapiforms

Oldest known fossil primates

K/T-Boundary

Inferred age of last common ancestor of living primates

Reconciling molecular and fossil records?

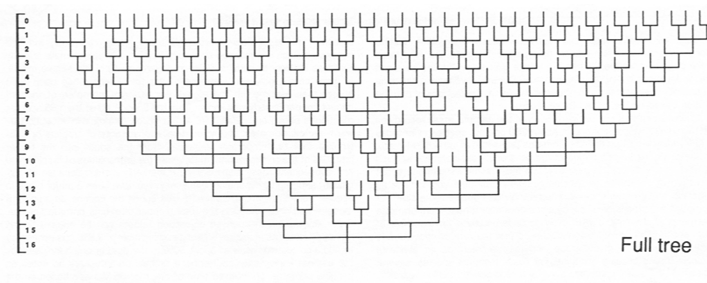
- Extant primates are strepsirrhines (lemurs and lorises) and haplorhines (tarsiers and anthropoids)
- Molecular estimate of time of divergence is approximately 90 mya
- Fossil record suggests 60-65 mya
- Fossil record is patchy

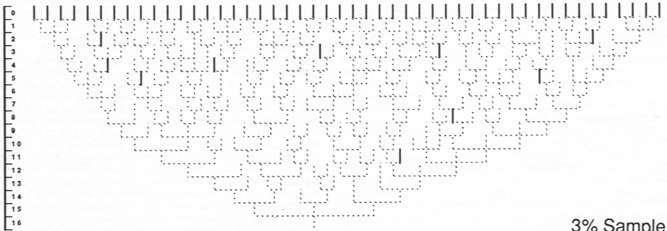
Reconciling molecular and fossil records?

- Extant primates are strepsirrhines (lemurs and lorises) and haplorhines (tarsiers and anthropoids)
- Molecular estimate of time of divergence is approximately 90 mya
- Fossil record suggests 60-65 mya
- Fossil record is patchy

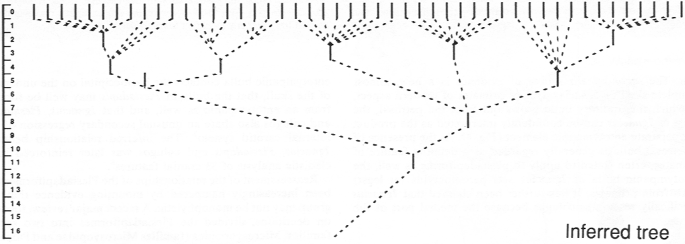
Problem: Use the fossil record to estimate the age of the last common ancestor of extant primates

Inference from the fossil record





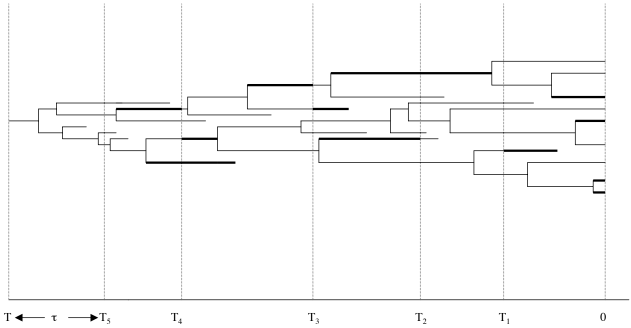
3% Sample


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Inferred tree

Primate Data

Epoch	k	T_k	Observed number of species (D_k)
Late Pleistocene	1	0.15	19
Middle Pleistocene	2	0.9	28
Early Pleistocene	3	1.8	22
Late Pliocene	4	3.6	47
Early Pliocene	5	5.3	11
Late Miocene	6	11.2	38
Middle Miocene	7	16.4	46
Early Miocene	8	23.8	36
Late Oligocene	9	28.5	4
Early Oligocene	10	33.7	20
Late Eocene	11	37.0	32
Middle Eocene	12	49.0	103
Early Eocene	13	54.8	68
Pre-Eocene	14		0

The evolutionary process



Statistical Aspects

- Highly dependent data
- Dependence caused by tree or graph linking observations (ancestral history)
- Explicit theory hard to come by . . .
- . . . computational inference methods essential

Stochastic computation in evolutionary genetics

- Many different approaches have been employed:
  - Rejection
  - Importance sampling
  - Sequential importance sampling with resampling
  - Markov chain Monte Carlo
  - Population Monte Carlo
  - Metropolis-coupled MCMC
  - Perfect sampling: coupling from the past
  - Variational Bayes methods

Rejection methods

Introduction to Bayesian computation

- Discrete data $\mathcal{D}$, prior $\pi(\theta)$ for parameters θ
- Want to generate observations from posterior distribution $f(\theta \mid \mathcal{D})$
- Bayes Theorem gives

$$f(\theta \mid \mathcal{D}) = \mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)/\mathbb{P}(\mathcal{D}),$$

where the normalizing constant is

$$\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D} \mid \tau)\pi(\tau)\, d\tau$$

Posterior proportional to likelihood × *prior*

Rejection methods I

R1 Generate θ from $\pi(\cdot)$

R2 Accept θ with probability $h = \mathbb{P}(\mathcal{D} \mid \theta)$; return to R1

- Accepted observations have distribution $f(\theta \mid \mathcal{D})$
- Can do better: if

$$\mathbb{P}(\mathcal{D} \mid \theta) \leq c \text{ for all } \theta$$

then can replace h above with h/c

The number of runs to get n observations from $f(\theta|\mathcal{D})$ is negative binomial, with mean

$$\frac{nc}{\mathbb{P}(\mathcal{D})}$$

This shows

- the effect of $\mathbb{P}(\mathcal{D})$
- the effect of c
- the way to estimate $\mathbb{P}(\mathcal{D})$ (Bayes factors)

Complex stochastic models

- A stochastic process often underlies the likelihood computation
- This process may be complex, making explicit probability calculations difficult or impossible
- Thus $\mathbb{P}(\mathcal{D} \mid \theta)$ may be uncomputable (either quickly or theoretically)

When the stochastic process is easy to simulate . . .

Rejection methods II

RS1 Generate θ from $\pi(\cdot)$

RS2 Simulate $\mathcal{D}'$ from stochastic model with parameter θ

RS3 Accept θ if $\mathcal{D}' = \mathcal{D}$; return to RS1

- Just as before, accepted observations from this algorithm have the density $f(\cdot|\mathcal{D})$
- Despite its appearance, this algorithm is much more general than first one — no need for explicit calculation
- Do we ever hit the target?

Approximate Bayesian Computation I

A1 Generate θ from $\pi(\cdot)$

A2 Simulate $\mathcal{D}'$ from stochastic model with parameter θ

A3 Calculate distance $\rho(\mathcal{D}, \mathcal{D}')$ between $\mathcal{D}'$ and $\mathcal{D}$

A4 Accept θ if $\rho \leq \epsilon$; return to 1

- If $\epsilon \to \infty$, generates from prior
- If $\epsilon = 0$, generates from $f(\theta \mid \mathcal{D})$

- Choice of ϵ reflects tension between computability and accuracy
  - PCR — post-computational remorse
- Method is *honest*: you get observations from $f(\theta \mid \rho(\mathcal{D}, \mathcal{D}') \leq \epsilon)$

An illustrative example (R. Wilkinson)

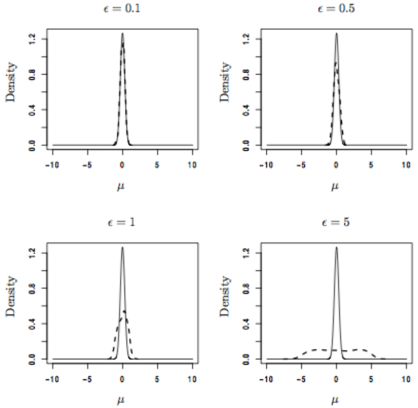
- $X_1, \ldots, X_n$ iid N(μ, σ^2)
- σ^2 known, improper prior $\pi(\mu) \propto 1$
- Posterior is N$(\bar{X}, \sigma^2/n)$
- Use ABC, accepting μ if $\bar{X}' \leq \epsilon$

  Can calculate:

$$\mathbb{E}(\mu||\bar{X}'| \leq \epsilon) = 0$$

$$\mathrm{Var}(\mu||\bar{X}'| \leq \epsilon) = \frac{\sigma^2}{n} + \frac{\epsilon^2}{3}$$

Plots of posterior for μ



Estimating the error

Can calculate

$$d_{TV}(\pi_\epsilon(\mu), \pi(\mu|\bar{X}=0) = \frac{1}{2}\int |\pi_\epsilon(\mu) - \pi(\mu|\bar{X}=0)|d\mu$$

Get

$$d_{TV} = \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2),$$

where $c = \sqrt{2/\pi}\exp(-1/2) \approx 1/2$.

Approximate Bayesian Computation II

- The limit $\epsilon = 0$ reproduces the data precisely
- In many examples the data are too high-dimensional
- . . . so reduce dimension by using summary statistics

Approximate sufficiency

Recall that $S = S(\mathcal{D})$ is *sufficient* for θ if

$$\mathbb{P}(\mathcal{D} \mid S, \theta) \text{ is independent of } \theta$$

- If S is sufficient for θ, then $f(\theta \mid \mathcal{D}) = f(\theta \mid S)$
- Typically, S is of smaller dimension that $\mathcal{D}$. Inference method can be simplified (and sped up), e.g. rejection method via

$$f(\theta \mid S) \propto \mathbb{P}(S \mid \theta)\pi(\theta)$$

Research problem

Puts a premium on finding decent summary statistics

- Definition of approximate sufficiency? (LeCam 1963)
- A systematic, implementable approach?
- Estimate distance between $f(\theta \mid \mathcal{D})$ and $f(\theta \mid S)$ given a measure of how far from sufficient S is for θ

Combine summaries and rejection

T et al (1997), Fu and Li (1997), Weiss and von Haeseler (1998), Pritchard et al (1999), Wall (2000)

Choose statistics $S = (S_1, \ldots, S_p)$ to summarize $\mathcal{D}$

```
AS1 Generate $\theta$ from $\pi(\cdot)$
AS2 Simulate $\mathcal{D}'$, calculate $s'$
AS3 Accept $\theta$ if $\rho(s', s) \leq \epsilon$; return to 1
```

ABC III: Generalization of ABC II

Beaumont, Zhang and Balding (2002), Genetics

A method that makes use of observations in a better way:

- No sharp cut-off
- Weight all simulated observations using distance from target

Advantages and disadvantages of ABC

Pros:

- Usually easy to code
- Generates independent observations (can use embarrassingly parallel computation)
- Can be used to estimate Bayes factors directly
- Usually easy to adapt

Cons:

- May be hard to anticipate effects of summary statistics
- For complex probability models, sampling from prior does not make good use of accepted observations
- Choice of metric matters
- Homework: find a scheme that combines generation of perfect observations with use of existing sample

Markov chain Monte Carlo

The Hastings Markov chain

M1 Now at θ

M2 Propose move to θ' according to $q(\theta \to \theta')$

M3 Calculate the Hastings ratio

$$h = \min\left(1, \frac{\mathbb{P}(\mathcal{D} \mid \theta')\pi(\theta')q(\theta' \to \theta)}{\mathbb{P}(\mathcal{D} \mid \theta)\pi(\theta)q(\theta \to \theta')}\right)$$

M4 Accept θ' with probability h, else return θ

Some special cases

- Independence sampler: $q(\theta, \theta') = g(\theta')$. h depends on ratios of likelihoods, priors and proposals

- Metropolis sampler: $q(\theta, \theta') = q(\theta', \theta)$. h depends on likelihood and ratio of priors

- Reversible sampler:

  $$\pi(\theta)q(\theta, \theta') = \pi(\theta')q(\theta', \theta)$$

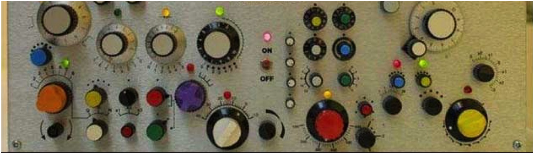
  h depends on likelihood ratio

Basic output analysis

There are more things to check:

- Is the chain ergodic?
- Does it mix well?
- Is the chain stationary?
- Burn in?
- Diagnostics of the run (no free lunches)

MCMC in evolutionary genetics setting



- Small tweaks in the biology often translate into huge changes in algorithm
- Long development time
- All the usual problems with convergence
- Almost all the effort goes into evaluation of likelihood

(Yet) another MCMC approach

MS1 Now at θ

MS2 Propose a move to θ' according to $q(\theta \to \theta')$

MS3 Generate $\mathcal{D}'$ using θ'

MS4 If $\mathcal{D}' = \mathcal{D}$, go to next step, else return θ

MS5 Calculate

$$h = h(\theta, \theta') = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right)$$

MS6 Accept θ' with probability h, else return θ

Practical version: ABC

Data $\mathcal{D}$, summary statistics S

[MS4$'$] If $\rho(\mathcal{D}', \mathcal{D}) \leq \epsilon$, go to next step, otherwise return θ

[MS4$''$] If $\rho(S', S) \leq \epsilon$, go to next step, otherwise return θ

for some suitable metric ρ and approximation level ϵ

Observations now from $f(\theta \mid \rho(\mathcal{D}', \mathcal{D}) \leq \epsilon)$ or $f(\theta \mid \rho(S', S) \leq \epsilon)$

Variations on ABC

- These methods can often be started at stationarity, so no burn-in
- If the underlying probability model is complex, simulating data will not often lead to acceptance. Thus need update for parts of the probability model (data augmentation)
- What if $\mathcal{D}$ is not discrete?
  - Use previous method (binning)
  - Use simulation approach to estimate the likelihood terms in the Hastings ratio (Diggle & Gratton, RSSB, 1980)

. . . the lecture ended here, but there is a bit more . . .

ABC-PRC

Partial rejection control – Sisson et al., PNAS, 2007

- Start with $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T = \epsilon$
- For iteration $t = 1, 2, \ldots, T$ the algorithm produces samples $(\theta_1^{(t)}, \ldots, \theta_N^{(t)})$

1. Set $t = 1$

2. Set $i = 1$

2A. If $t = 1$, sample $\theta^{**} \sim \mu_1$

If $t > 1$, sample θ^* from $\{(\theta_{t-1}^{(j)}, w_{t-1}^{(j)})\}$

Generate $\theta^{**} \sim K_t(\cdot|\theta^*)$

Generate data $\mathcal{D}^{**}$ from model with parameter θ^{**}

If $\rho(S(\mathcal{D}^{**}), S(\mathcal{D})) \geq \epsilon_t$, go to 2A

2B. Set

$$\theta_t^{(i)} = \theta^{**}, \quad w_t^{(i)} = \frac{\pi(\theta_t^{(i)}) L_{t-1}(\theta^*|\theta_t^{(i)})}{\pi(\theta^*) K_t(\theta_t^{(i)}|\theta^*)}$$

(with $w_1^{(i)} = \pi(\theta_1^{(i)})/\mu_1(\theta^{(i)})$

If $i < N$, set $i = i + 1$ and go to 2A

3. Normalize weights so that $\sum_{i=1}^{N} w_i^{(t)} = 1$.

   [resampling step to generate new sample $\{(\theta_t^{(j)}, w_t^{(j)} = 1/N)\}$]

4. If $t < T$, set $t = t + 1$ and go to 2.

The suggestion is to take $L_{t-1}(\theta'|\theta) = K_t(\theta|\theta')$ and $\mu_1 = \pi$, which removes all the weights!

3. Normalize weights so that $\sum_{i=1}^{N} w_i^{(t)} = 1$.

   [resampling step to generate new sample $\{(\theta_t^{(j)}, w_t^{(j)} = 1/N)\}$]

4. If $t < T$, set $t = t + 1$ and go to 2.

The suggestion is to take $L_{t-1}(\theta'|\theta) = K_t(\theta|\theta')$ and $\mu_1 = \pi$, which removes all the weights

. . . but turns out that this is biased!

ABC-PMC

Population Monte Carlo – Beaumont et al., Biometrika, 2008

Note that the t-th sample is produced from the proposal distribution

$$\hat{\pi}_t(\theta^{(t)}) \propto \sum_{j=1}^{N} w_j^{(t-1)} K_t(\theta^{(t)} | \theta_j^{(t-1)})$$

This suggests using the weights

$$w_i^{(t)} \propto \pi(\theta_t^{(i)}) \Big/ \hat{\pi}_t(\theta_i^{(t)})$$

- Furthermore, it is known that the kernel K_t MUST be changed in each iteration to get better accuracy
- Beaumont et al. have an optimised Gaussian updating scheme to do this

AWF Edwards. Biometrics, 23, p176, 1967

“It will be maintained that the end of an era has now been reached, as regards both statistical methods and computational techniques, and an outline of the way in which biometric techniques in genetical demography may be expected to develop will be given. Particular emphasis will be placed on the need to formulate sound methods of ‘estimation by simulation’ on complex models”.

Acknowledgments

- Rich Wilkinson (Maths, University of Sheffield)
- Paul Joyce (Maths, University of Idaho)
- Paul Marjoram, Peter Calabrese (University of Southern California)

References

Marjoram & Tavaré (2006) *Nat Rev Genet*, **7**, 759–770

Marjoram et al (2003) *Proc Natl Acad Sci USA*, **100**, 15324–15328

Nordborg & Tavaré (2002) *Trends in Genetics*, **18**, 83–90

Tavaré et al (2002) *Nature*, **416**, 726–729

Back to inference in the fossil record

ABC Approach

Data can be thought of in two parts:

(a) the observed number of fossils F_{obs} found

(b) the proportions $p_{j,\mathrm{obs}}$ found in jth bin

A suitable metric might be

$$\left|\frac{F}{F_{\mathrm{obs}}} - 1\right| + \frac{1}{2}\sum_{j=1}^{k+1} |p_j - p_{j,\mathrm{obs}}|$$

Note: no data summaries here

Sensitivity: Exploring Other Models

One advantage of ABC – it is easy to change the input . . .

- Choice of d
- Demography
- Sampling fractions
- K/T crash 65 mya
    - the time of origin of primates is even further back in the Cretaceous
- Poisson sampling scheme: length in bin matters
- Dating other split points

Modeling

- Diversification model: non-homogeneous Markov branching process
  - parameters: $\boldsymbol{\lambda}$, τ
- Sampling model: binomial sampling
  - parameters: $\boldsymbol{\alpha}$
  - $\boldsymbol{\alpha} = \alpha\boldsymbol{p}, \boldsymbol{p}$ known

Prior for Sampling Fractions

$$f(\boldsymbol{\lambda}, \tau, \mathcal{N}, \boldsymbol{\alpha} | \mathcal{D}) \propto \mathbb{P}(\mathcal{D} | \boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau, \mathcal{N}) \mathbb{P}(\mathcal{N} | \tau, \boldsymbol{\lambda}) f(\tau) f(\boldsymbol{\lambda}) f(\boldsymbol{\alpha})$$

where

- $\boldsymbol{\lambda} = (\lambda, \gamma, \rho)$ growth parameters
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{14})$ sampling fractions
- $\mathcal{N}$ is the underlying tree structure

Give sampling fractions independent Beta(a, b) priors

Gibbs-ABC Example

Rich Wilkinson (Sheffield)

Split the random variable into two parts:
$\boldsymbol{\alpha}$ and $(\boldsymbol{\lambda}, \tau, \mathcal{N})$

Sample from the two conditional distributions

- $f(\boldsymbol{\alpha} \mid \mathcal{D}, \boldsymbol{\lambda}, \tau, \mathcal{N})$
  - independent beta components
  - mean of $\alpha_i = \frac{a+d_i}{N_i+a+b} \approx \frac{d_i}{N_i}$
- $f(\tau, \boldsymbol{\lambda}, \mathcal{N} \mid \mathcal{D}, \boldsymbol{\alpha})$

Conditional distribution of $(\tau, \boldsymbol{\lambda}, \mathcal{N})$

$$\begin{aligned} f(\tau, \boldsymbol{\lambda}, \mathcal{N} | \mathcal{D}, \boldsymbol{\alpha}) &\propto f(\boldsymbol{\lambda}, \tau, \mathcal{N}, \boldsymbol{\alpha} | \mathcal{D}) \\ &\propto \mathbb{P}(\mathcal{D} | \boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathcal{N}, \alpha) \mathbb{P}(\mathcal{N} | \tau, \lambda) f(\tau) f(\boldsymbol{\lambda}) \end{aligned}$$

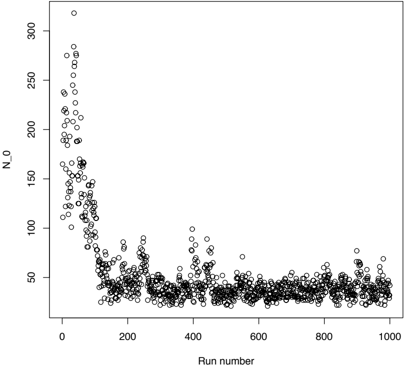
Simulate from this using ABC: accept $(\boldsymbol{\lambda}, \tau, \mathcal{N})$ if $\rho(\mathcal{D}, \mathcal{D}') < \epsilon$, where $\mathcal{D}'$ represents the simulated data

Metric and Priors

$$\begin{aligned}
\tau &\sim U[0,100] \\
\alpha &\sim U[0,0.6] \\
\rho &\sim U[0,0.8] \\
\gamma &\sim U[0.005,0.015] \\
1/\lambda &\sim U[2,3] \\
a &= 0.1 \\
b &= 1 \\
\epsilon &= 0.2
\end{aligned}$$

Same metric as before

Not going so well . . .



Tweak metric

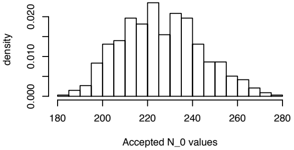
- The observed N_0 values are too small
    - require $N_0 > 235$
    - change the metric

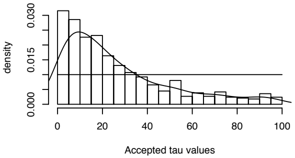
$$\rho(\mathcal{D}, \mathcal{D}') = \sum_{i=1}^{k} \left| \frac{D_i}{D_+} - \frac{D_i'}{D_+'} \right| + \left| \frac{D_+'}{D_+} - 1 \right| + \left| \frac{N_0'}{N_0} - 1 \right|$$

- Penalises trees with N_0 values far from 235

Results: $\epsilon = 0.3$

	min	LQ	Median	mean	UQ	Max
N_0	184	212	224	226	238	279
τ	0.0	8.0	18.6	26.3	36.8	99.5


density
Accepted N_0 values


density
Accepted tau values