

Species Tree Inference

Bruce Rannala @ UC Davis

Outline

Coalescent Theory

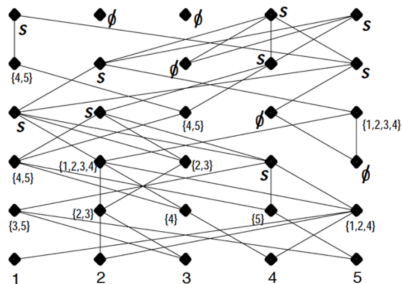
Multispecies Coalescent

Phyloclcmographic Inference

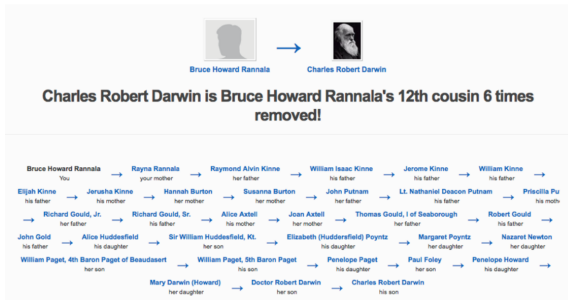
Species Tree Inference

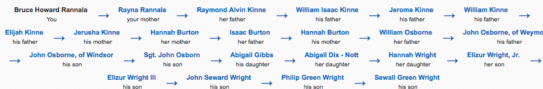
Species Delimitation

Pedigree

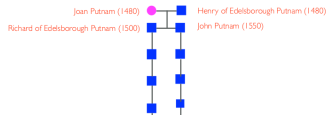


Pedigree



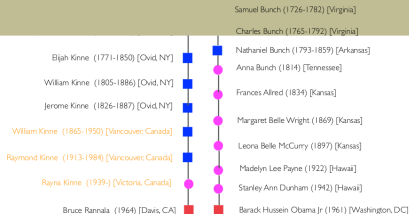


Pedigree

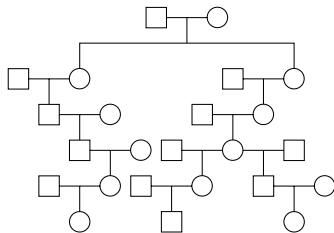


Kinship coefficient of Bruce Rannala and Barack Obama:

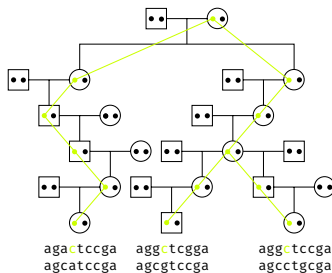
$$(1/2)^{34} = 5.8 \times 10^{(-11)}$$



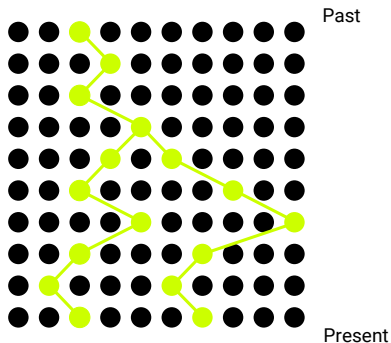
Pedigree



Gene Tree Within a Pedigree

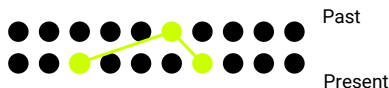


Population Coalescent ($n = 2$)

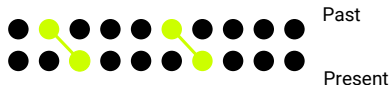


Population Coalescent ($n = 2$)

Probability of common parent ($P_1 = \frac{1}{2N}$):



Probability of distinct parents ($P_2 = 1 - \frac{1}{2N}$):



Population Coalescent ($n = 2$)

Probability no coalescence by generation t

$$P_2^{(t)} = \left(1 - \frac{1}{2N}\right)^t$$

Transform to “diffusion” timescale

$$t = (2N)\tau$$

Unit of time is now $2N$ generations. One generation on this timescale is

$$d\tau = \frac{1}{2N}$$

so as $N \rightarrow \infty$ time appears continuous.

Population Coalescent ($n = 2$)

Probability of no coalescence (on the transformed timescale) is

$$P_2^{(\tau)} = \left(1 - \frac{1}{2N}\right)^{\tau(2N)}$$

and for large population size this converges to

$$\lim_{N \rightarrow \infty} \left(1 - \frac{1}{2N}\right)^{\tau(2N)} = e^{-\tau}$$

Expectation of Coalescence Time ($n = 2$)

On the original discrete generation timescale the expected time for 2 sequences to coalesce is

$$\begin{aligned}\mathbb{E}(t) &= \sum_{t=0}^{\infty} t P_2^{(t)} \\ &= \sum_{t=1}^{\infty} t \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} = 2N.\end{aligned}$$

On the transformed continuous timescale the expected time to coalescence is

$$\mathbb{E}(\tau) = \int \tau P_2^{(\tau)} d\tau = \int_0^{\infty} \tau e^{-\tau} d\tau = 1.$$

Distribution of TMRCA ($n = 2$)

Recall: an exponentially distributed random variable x with rate λ has probability density function

$$f(x) = \lambda e^{-\lambda x},$$

with mean (expectation)

$$\mathbb{E}(x) = 1/\lambda,$$

and variance

$$\text{Var}(x) = 1/(\lambda^2).$$

Distribution of TMRCA ($n = 2$)

The probability density function of the coalescence time for two sequences on the transformed timescale ($2N$ generations) is

$$f(\tau) = e^{-\tau},$$

which is an exponential distribution with $\lambda = 1$. The mean and variance are

$$\mathbb{E}(\tau) = 1,$$

and

$$\text{Var}(\tau) = 1.$$

Distribution of TMRCA ($n = 2$)

The mean and variance of the coalescence time for two sequences on the original timescale (generations) are

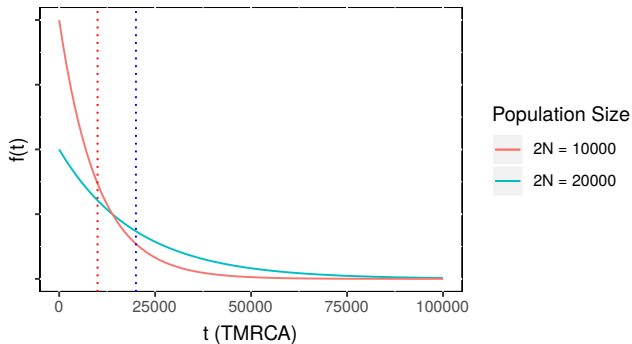
$$\begin{aligned}\mathbb{E}(t) &= \mathbb{E}[(2N)\tau] \\ &= (2N)\mathbb{E}(\tau) \\ &= 2N,\end{aligned}$$

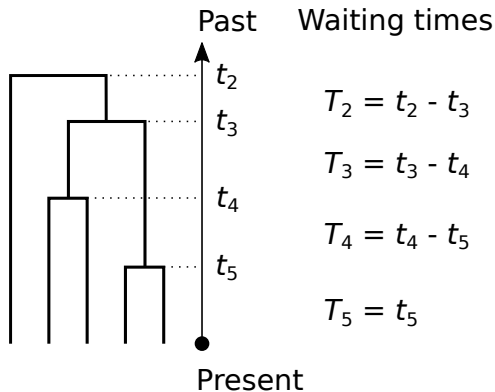
and

$$\begin{aligned}\text{Var}(t) &= \text{Var}[(2N)\tau] \\ &= (2N)^2\text{Var}(\tau) \\ &= 4N^2.\end{aligned}$$

Distribution of TMRCA ($n = 2$)

Probability density of coalescence time ($n=2$)



Population Coalescent ($n \geq 2$)

Population Coalescent ($n \geq 2$)

Waiting time for i lineages to coalesce to $i - 1$ lineages

$$f(\tau_i) = \frac{i(i-1)}{2} e^{-\frac{i(i-1)}{2}\tau_i}$$

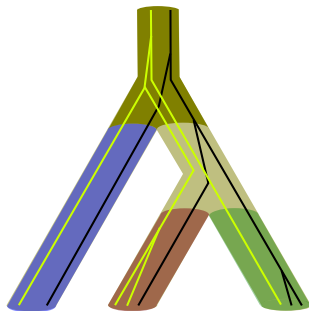
This is an exponential distribution with rate parameter

$$\frac{i(i-1)}{2}$$

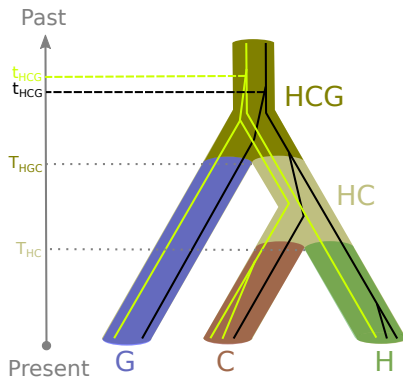
The mean and variance are therefore

$$\mathbb{E}(\tau_i) = \frac{2}{i(i-1)}, \quad \text{Var}(\tau_i) = \frac{4}{i^2(i-1)^2}.$$

Gene Trees Within Species Trees

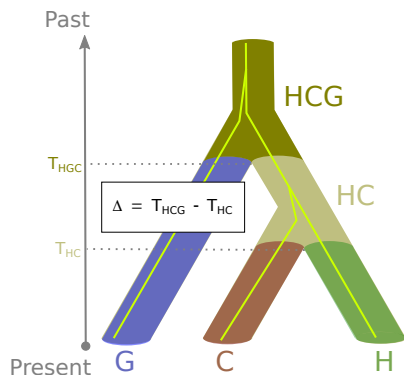


Gene Trees Within Species Trees



Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each



Probability H and C
coalesce in HC

$$\int_0^{\Delta} \frac{e^{-\frac{x}{2N_{HC}}}}{2N_{HC}} dx = 1 - e^{-\frac{\Delta}{2N_{HC}}}.$$

Probability H and C do not
coalesce in HC

$$e^{-\frac{\Delta}{2N_{HC}}}.$$

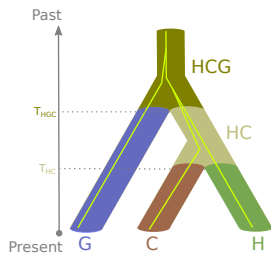
Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each

Species Tree = Gene Tree

$$\text{Pr} = 1/3$$

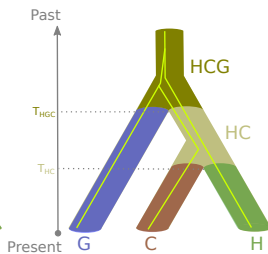
$$1/3$$



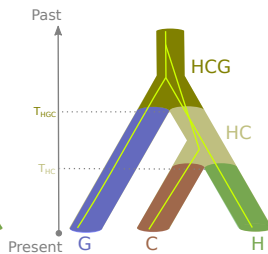
Species Tree \neq Gene Tree

$$\text{Pr} = 1/3 + 1/3 = 2/3$$

$$1/3$$



$$1/3$$



Gene Tree Probabilities Within Species Trees

3 species: 1 sequence each

$$\begin{aligned}\Pr(G = S) &= 1 - \exp\left(\frac{-\Delta}{2N_{HC}}\right) + \frac{1}{3} \exp\left(\frac{-\Delta}{2N_{HC}}\right) \\ &= 1 - \frac{2}{3} \exp\left(\frac{-\Delta}{2N_{HC}}\right)\end{aligned}$$

$$\Pr(G \neq S) = \frac{2}{3} \exp\left(\frac{-\Delta}{2N_{HC}}\right)$$

Estimator of Ancestral N (Chen and Li, 2001)

Procedure: estimate gene trees from sequence data and check match with a known species tree. The expected proportion of matches is

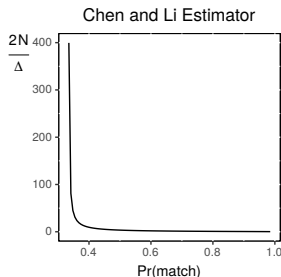
$$P = 1 - \frac{2}{3} \exp\left(\frac{-\Delta}{2N}\right)$$

Solving for $2N/\Delta$ gives the estimator:

$$\frac{2N}{\Delta} = \frac{1}{-\log(3/2) - \log(1 - P)}.$$

Estimator of Ancestral N (Chen and Li, 2001)

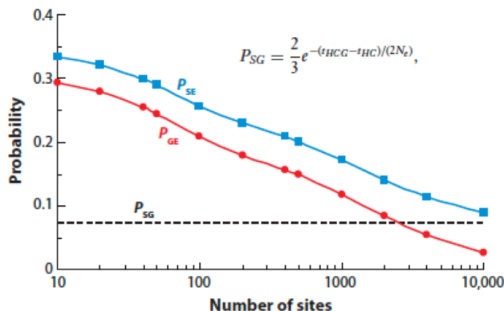
Population Size Versus Match Probability



Example: Let $\Delta_{HC} = 4 \times 10^6$ years (2×10^5 generations if $g = 20$) and $N_{HC} = 10^5$.

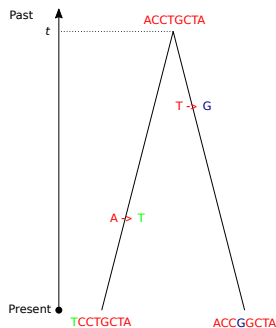
$$\begin{aligned} P &= 1 - \frac{2}{3} \exp\left(-\frac{200000}{2 \times 100000}\right) \\ &= 1 - \frac{2}{3} \exp(-1) \\ &= 1 - \frac{2}{3} \times 0.368 = 0.755 \end{aligned}$$

Estimator of Ancestral N (Chen and Li, 2001)



What is θ ?

two sequences



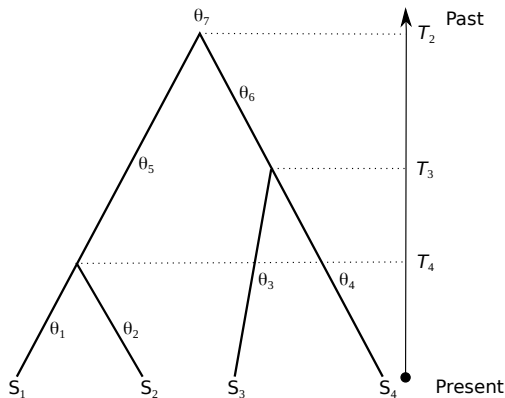
Number of segregating sites under infinite-sites model (Watterson, 1981)

$$\begin{aligned}\mathbb{E}(S) &= \mathbb{E}(2\mu t) \\ &= 2\mu \mathbb{E}(t) \\ &= 2\mu(2N) \\ &= 4N\mu\end{aligned}$$

We define

$$\theta = 4N\mu$$

Parameters of Phylogenetic Model



Bayesian Phylodemographic Inference

Let $\Theta = \{\theta, \tau\}$. For s species, θ contains at most $2s - 3$ and at least $s - 1$ parameters. τ contains $s - 1$ parameters. The posterior distribution of demographic parameters given sequence data D is

$$f(\Theta|D) = \int \frac{f(D|G)f(G|\Theta)f(\Theta)}{f(D)} dG$$

where $f(D|G)$ is the “Felsenstein Likelihood” and $f(G|\Theta)$ is the “Multispecies Coalescent” prior on gene trees.

Bayesian Phylogenetic Inference

Assumptions

Unlinked genes

$$f(G|\Theta) = \prod_{i=1}^L f(G_i|\Theta)$$

Independent sites

$$f(D|G_i) = \prod_{i=1}^n f(D_i|G_i)$$

No gene flow between populations.

Bayesian Phylogenetic Inference

Markov chain Monte Carlo

1. Simulate a proposed value for a parameter

$$\theta^* \approx g(\theta^*|\theta)$$

2. Accept proposed value with probability

$$\alpha = \min \left(\frac{f(D|\theta^*)f(\theta^*)g(\theta|\theta^*)}{f(D|\theta)f(\theta)g(\theta^*|\theta)}, 1 \right).$$

Metropolis et al. (1953) + Hastings (1970)

Bayesian Phylogenetic Inference

Metropolis-Hastings Algorithm

Edward Teller (father of the H-bomb)



Bayesian Phylogenetic Inference

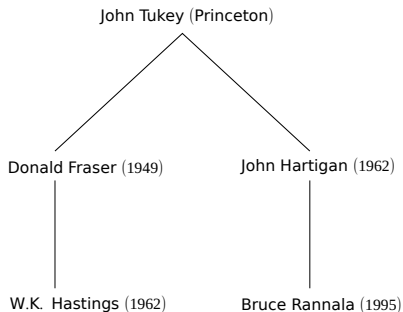
Metropolis-Hastings Algorithm

Keith Hastings (statistician)



Bayesian Phylogenetic Inference

Metropolis-Hastings Algorithm



Bayesian Phylogenetic Inference

MCMC Proposal Moves

- ▶ Propose changes to coalescent times in gene trees that respect constraints of species tree
- ▶ Propose changes to gene trees by subtree pruning and re-grafting (respect species tree constraints)
- ▶ Propose changes to effective population size parameters
- ▶ Propose new speciation times in the species tree and transforming gene trees to respect constraints
- ▶ Jointly propose proportional changes to all effective population sizes, divergence, and coalescence times

Bayesian Phylogenetic Inference

Making Sense of BPP Parameters

$$\theta = 4N\mu$$

units are expected DNA substitutions.

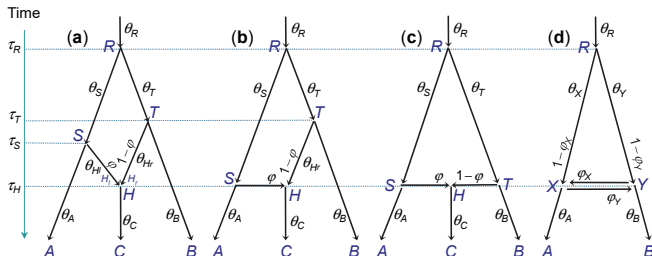
To obtain N we specify a mutation rate and generation time

$$N = \frac{\theta}{4 \times \mu \times g}$$

Example: if $\theta_H = 0.00057$, $g = 20$ years/generation and $\mu = 10^{-9}$ mutations/year

$$N_H = \frac{0.00057}{4 \times 10^{-9} \times 20} = 7125$$

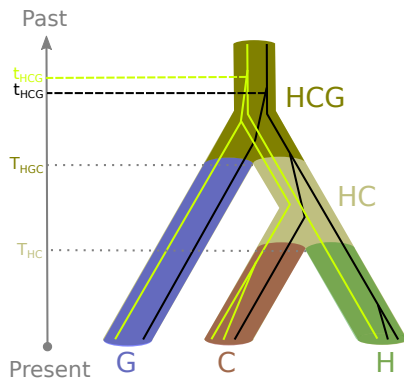
Multispecies Coalescent With Introgression



Multispecies Coalescent With Introgression

Standard Newick notation such as $(A(B,C))$ does not work for trees with introgression. Extended Newick format is very confusing. An explanation will be attempted by Professor Yang.

Infer Species Tree Topology and Parameters



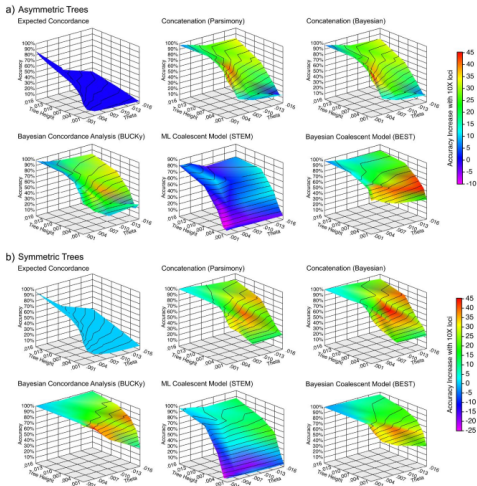
Approaches to Species Tree Inference

Concatenation: assumes that all gene trees have the same topology and branch lengths.

Alternative: allow different gene trees. Parametric methods model processes influencing the distribution of gene trees.

- ▶ multi-species coalescent process
- ▶ ongoing or periodic gene flow
- ▶ hybrid species
- ▶ recombination
- ▶ gene duplication (or loss)
- ▶ selection

Performance of Species Tree Methods

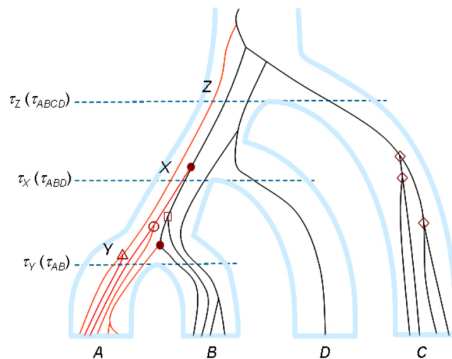


Infer Species Tree Topology and Parameters

Let S be the species tree topology and $\Theta = \{\theta, \tau\}$. The posterior probability of the species tree is

$$f(S|D) = \int_G \frac{f(D|\Theta, G)f(G|\Theta, S)f(\Theta)f(S)}{f(D)} dG$$

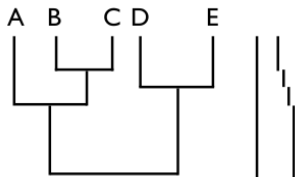
Species Tree Proposals



- Species Tree Inference

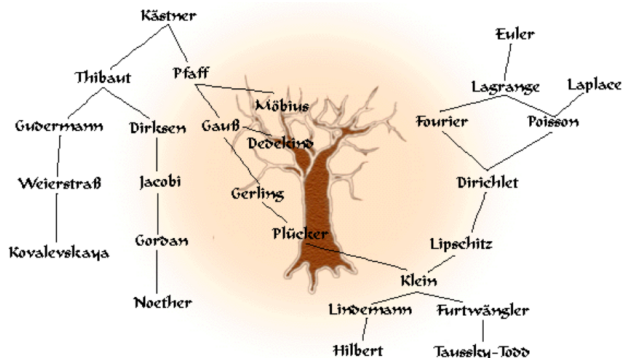
Prior on branch lengths

- Yule or Birth-Death Process (*Beast)
- Dirichlet Distribution conditioned on root age (BPP)



Prior on Species Trees

Mathematics Genealogy Project



Quick Search

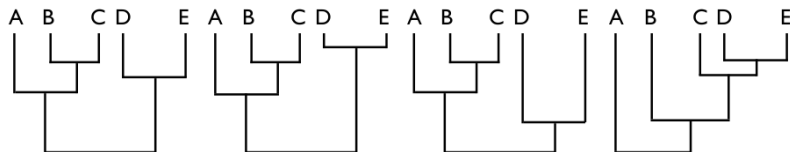
Search

[Advanced Search](#)

Prior on Species Trees

Prior on topology

- ▶ Uniform on labelled histories (*Beast, BPP)
- ▶ Uniform on labelled trees



speciesmodelprior = 0 (uniform LH)

speciesmodelprior = 1 (uniform rooted trees)

Placeholder

A placeholder