Coalescent Likelihood Methods

Mary K. Kuhner
Genome Sciences
University of Washington
Seattle WA

Outline

1. **Introduction to coalescent theory**
2. **Practical example**
3. **Genealogy samplers**
4. **Break**
5. **Survey of samplers**
6. **Evolutionary forces**
7. **Practical considerations**

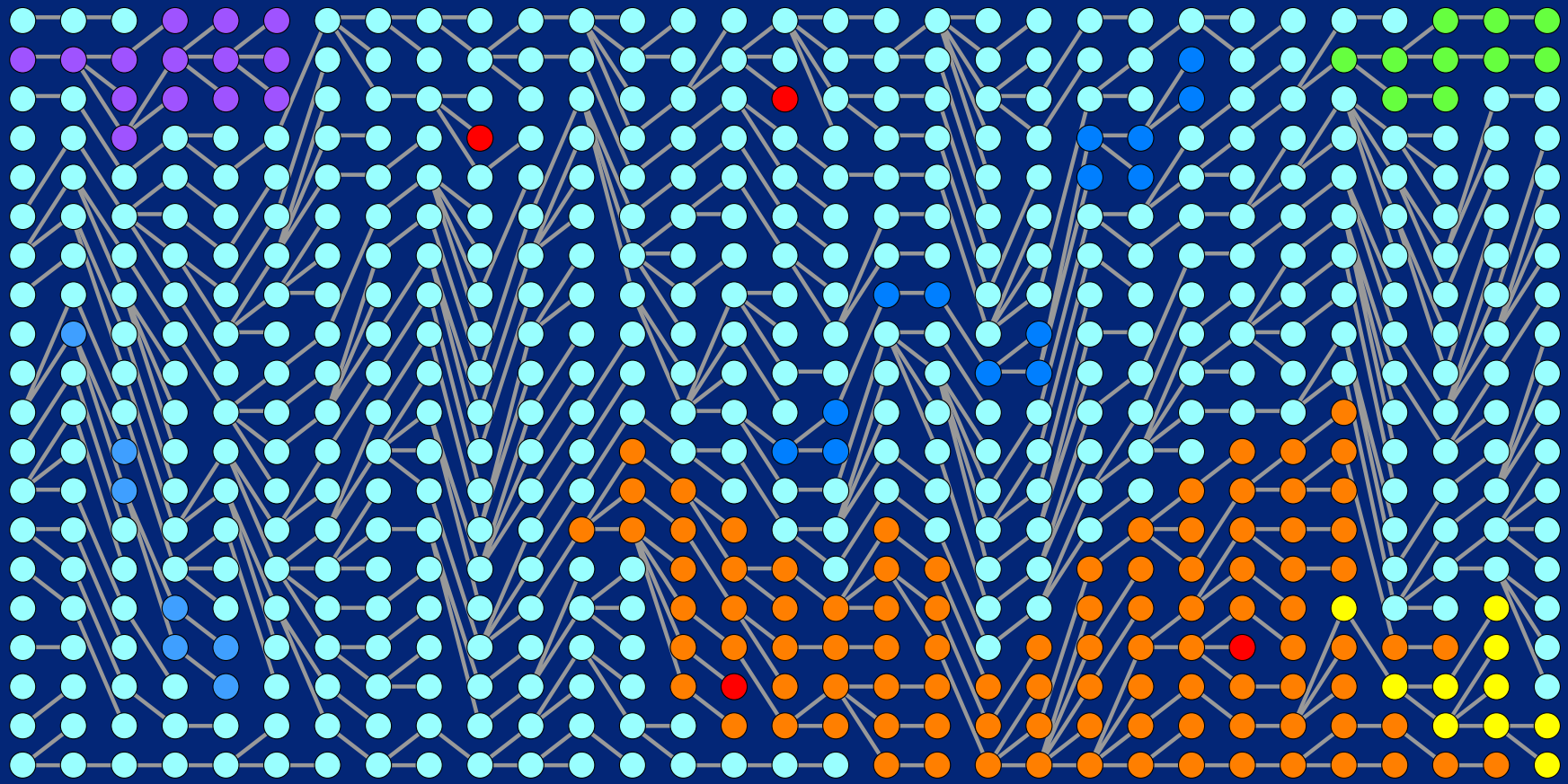
Population genetics can help us to find answers

- We are interested in questions like
  - How big is this population?
  - Are these populations isolated? How common is migration?
  - How fast have they been growing or shrinking?
  - What is the recombination rate across this region?
  - Is this locus under selection?
- All of these questions require comparison of many individuals.

Coalescent-based studies

- How many gray whales were there prior to whaling?
- When was the common ancestor of HIV lines in a Libyan hospital?
- Is the highland/lowland distinction in Andean ducks recent or ancient?
- Did humans wipe out the Beringian bison population?
- What proportion of HIV virions in a patient actually contribute to the breeding pool?
- What is the direction of gene flow between European rabbit populations?

Basics: Wright-Fisher population model



All individuals release many gametes and new individuals for the next generation are formed randomly from these.

Wright-Fisher population model

- Population size N is constant through time.
- Each individual gets replaced every generation.
- Next generation is drawn randomly from a large gamete pool.
- Only genetic drift affects the allele frequencies.

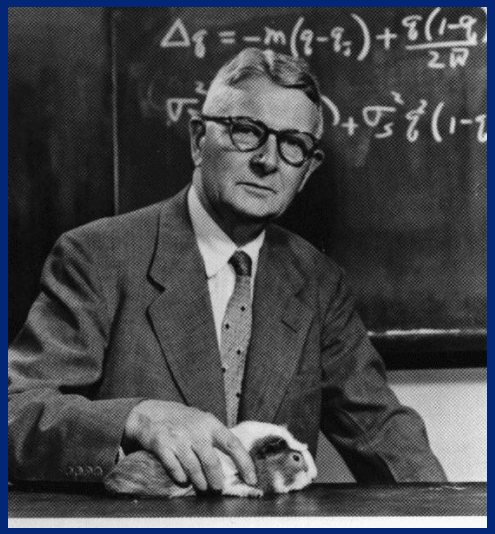
Other population models

- Other population models can often be equated to Wright-Fisher
- The N parameter becomes the effective population size N_e
- For example, cyclic populations have an N_e that is the harmonic mean of the various sizes

The big trick

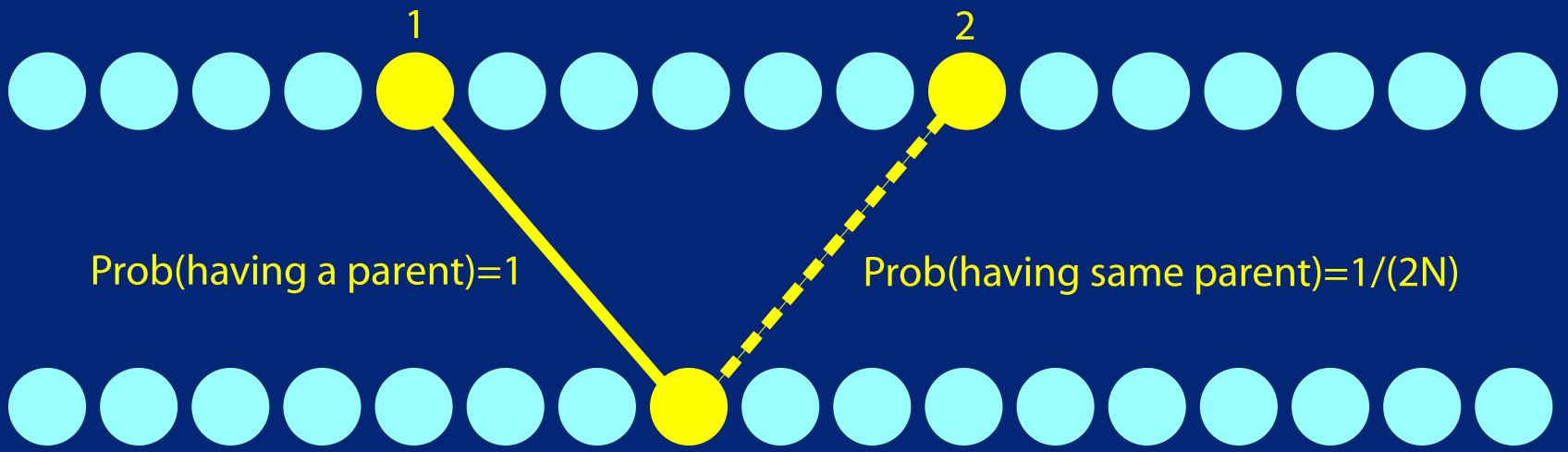
- We have a model for the progress of a population forward in time
- What we observe is the end product: genetic data today
- We want to reverse this model so that it tells us about the $past$ of our sequences

The Coalescent

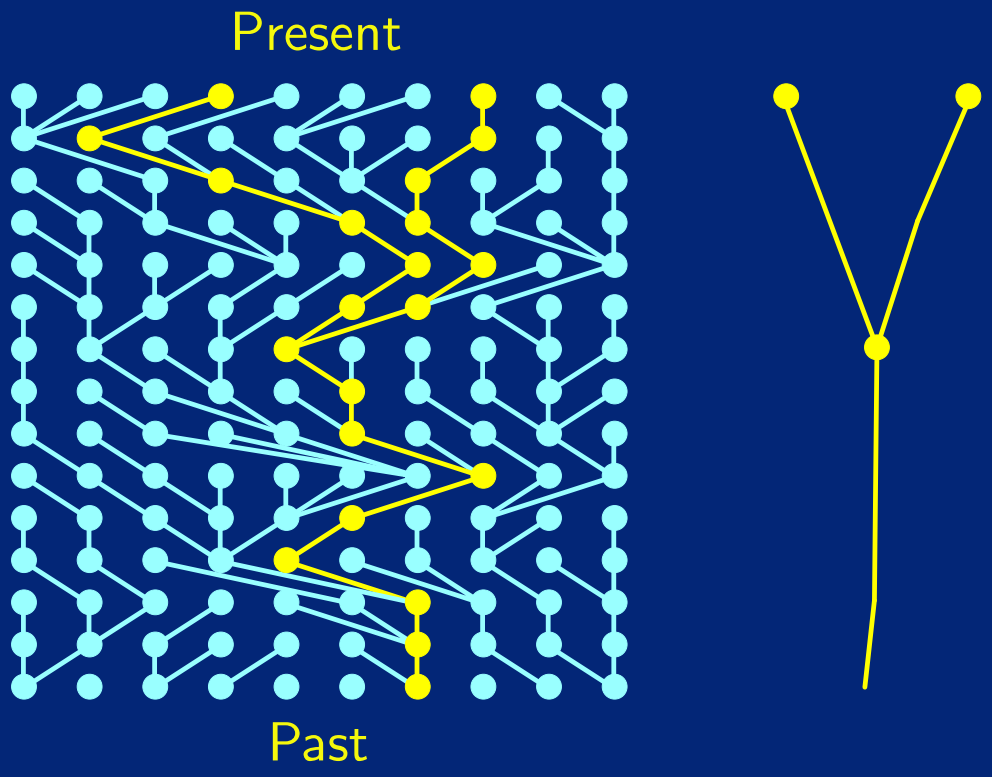


Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation is

$$\text{Prob (two genes share a parent)} = \frac{1}{2N}$$



The Coalescent



In every generation, there is a chance of $1/2N$ to coalesce. Following the sampled lineages through generations backwards in time we realize that it follows a geometric distribution with

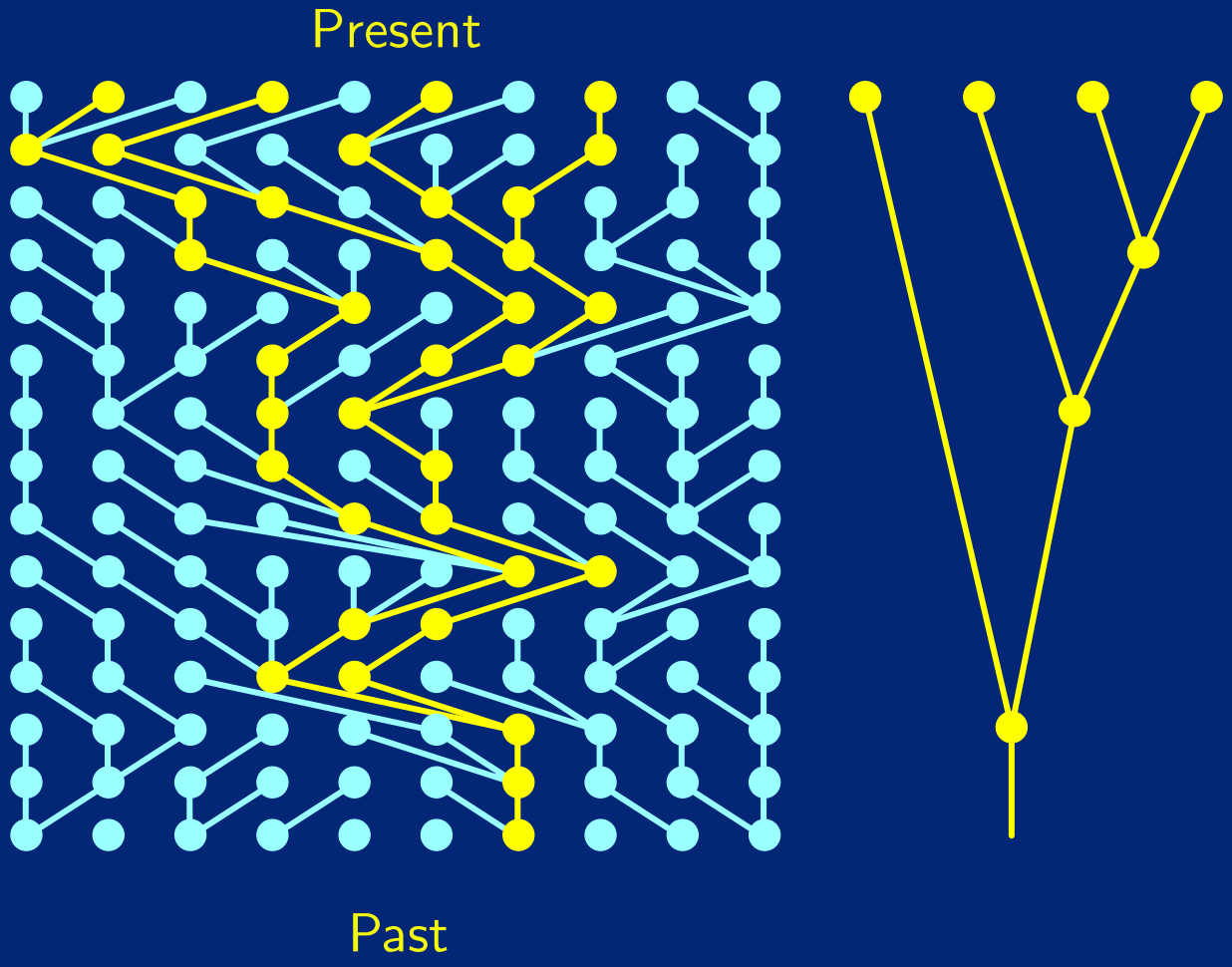
$\mathbb{E}(u) = 2N$ [the expectation of the time of coalescence u of **two** tips is $2N$]

The Coalescent

JFC Kingman generalized this for k gene copies.

$$\text{Prob } (k \text{ copies are reduced to } k-1 \text{ copies}) = \frac{k(k-1)}{4N}$$

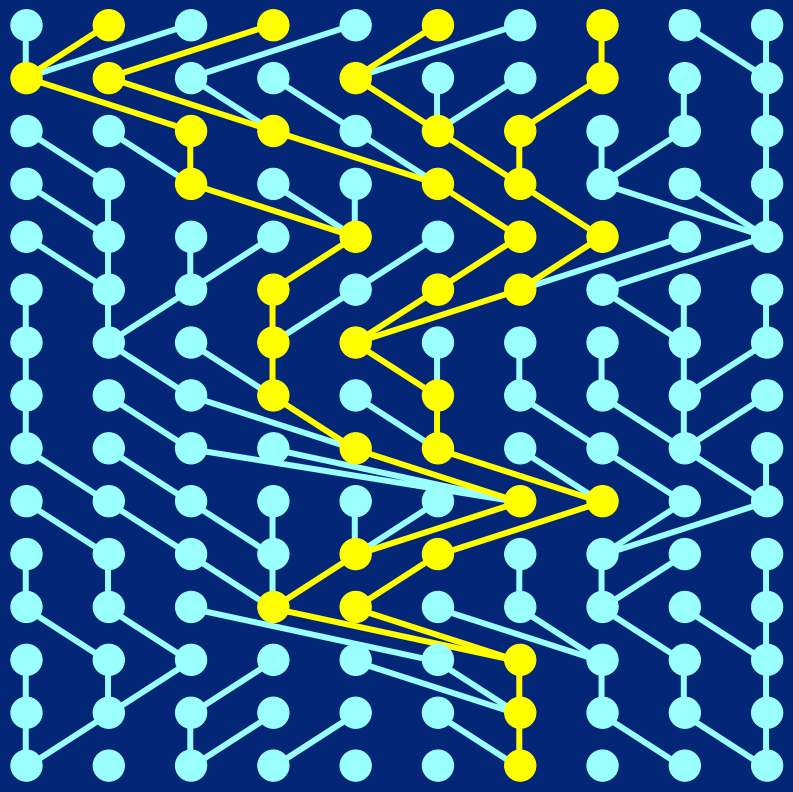
Kingman's n-coalescent

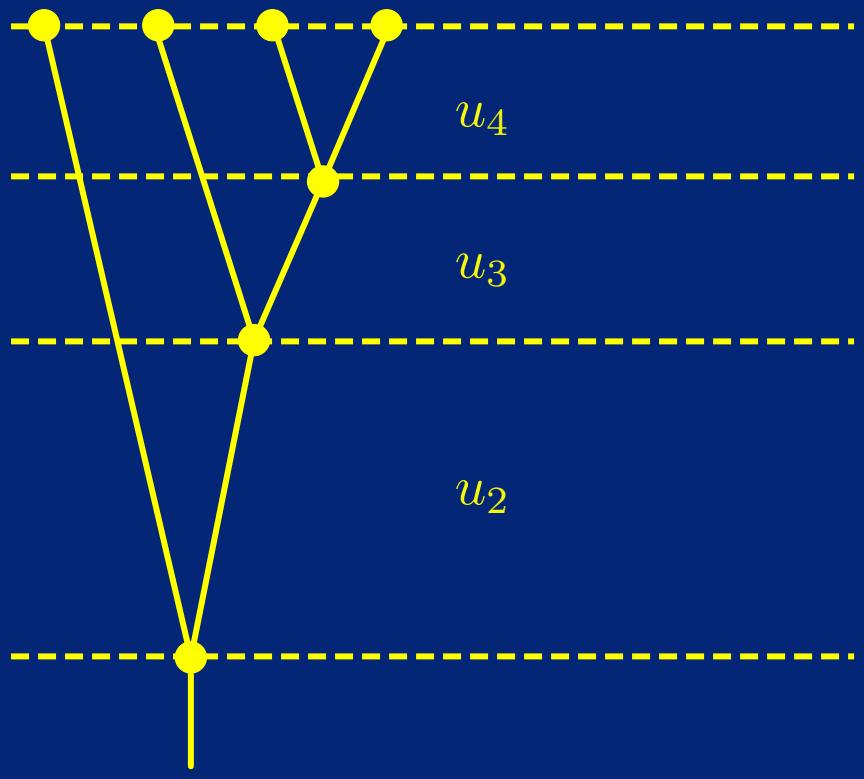


Kingman's n-coalescent

The expectation for the time interval u_k is

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)}$$





$$\mathrm{p(G|N)} = \prod_{\mathrm{i}} \exp(-\mathrm{u_i}\frac{\mathrm{k(k-1)}}{\mathrm{4N}})\frac{1}{\mathrm{2N}}$$

The Θ parameter

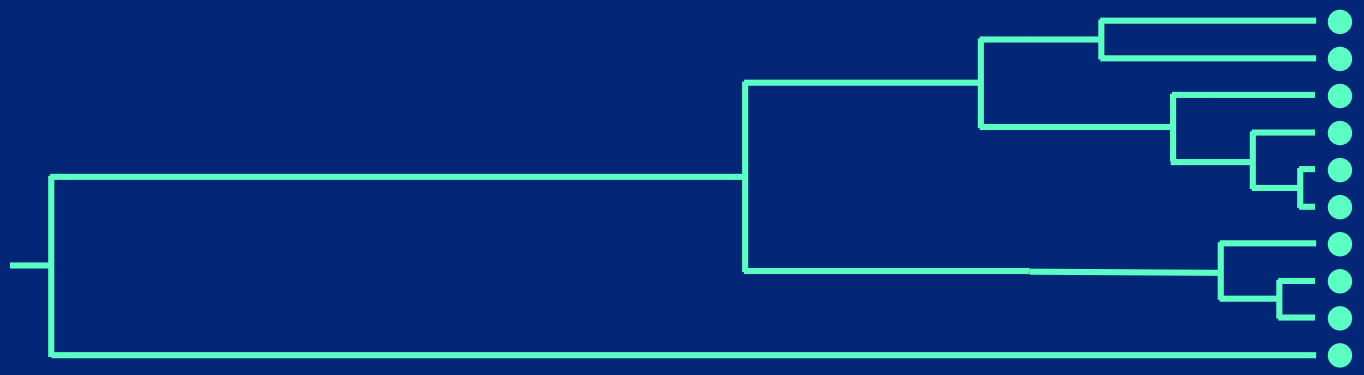
- The n-coalescent is defined in terms of N_e and time.
- We cannot measure time just by looking at genes, though we can measure divergence.
- We rescale the equations in terms of N_e, time, and the mutation rate μ.
- We can no longer estimate N_e but only the composite parameter Θ.
- $\Theta = 4N_e\mu$ in diploids.
- Multiple time point data can separate N_e and μ

What is this coalescent thing good for?

Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

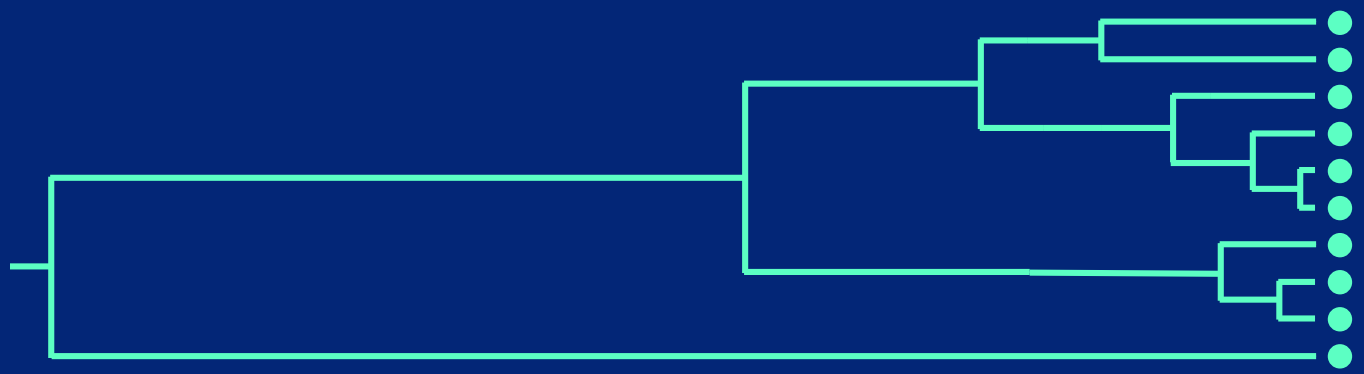
2. We know that we can calculate $\mathrm{p}(\text{Genealogy}|\mathrm{N})$



Utopian population size estimator

1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

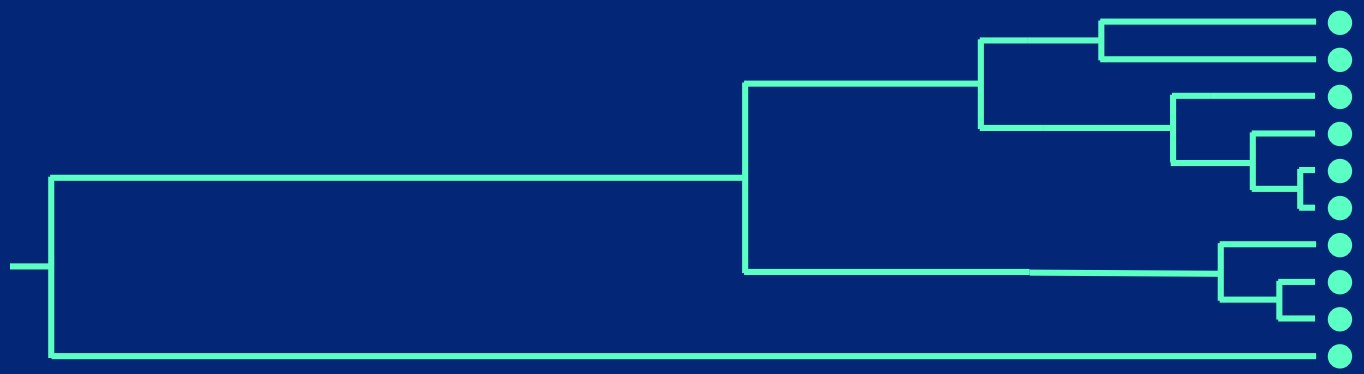


$$\mathrm{p(G|N)} = \mathrm{p(u_1|N,k)}\frac{1}{\mathrm{2N}} \times \mathrm{p(u_2|N,k-1)}\frac{1}{\mathrm{2N}} \times .....$$

Utopian population size estimator

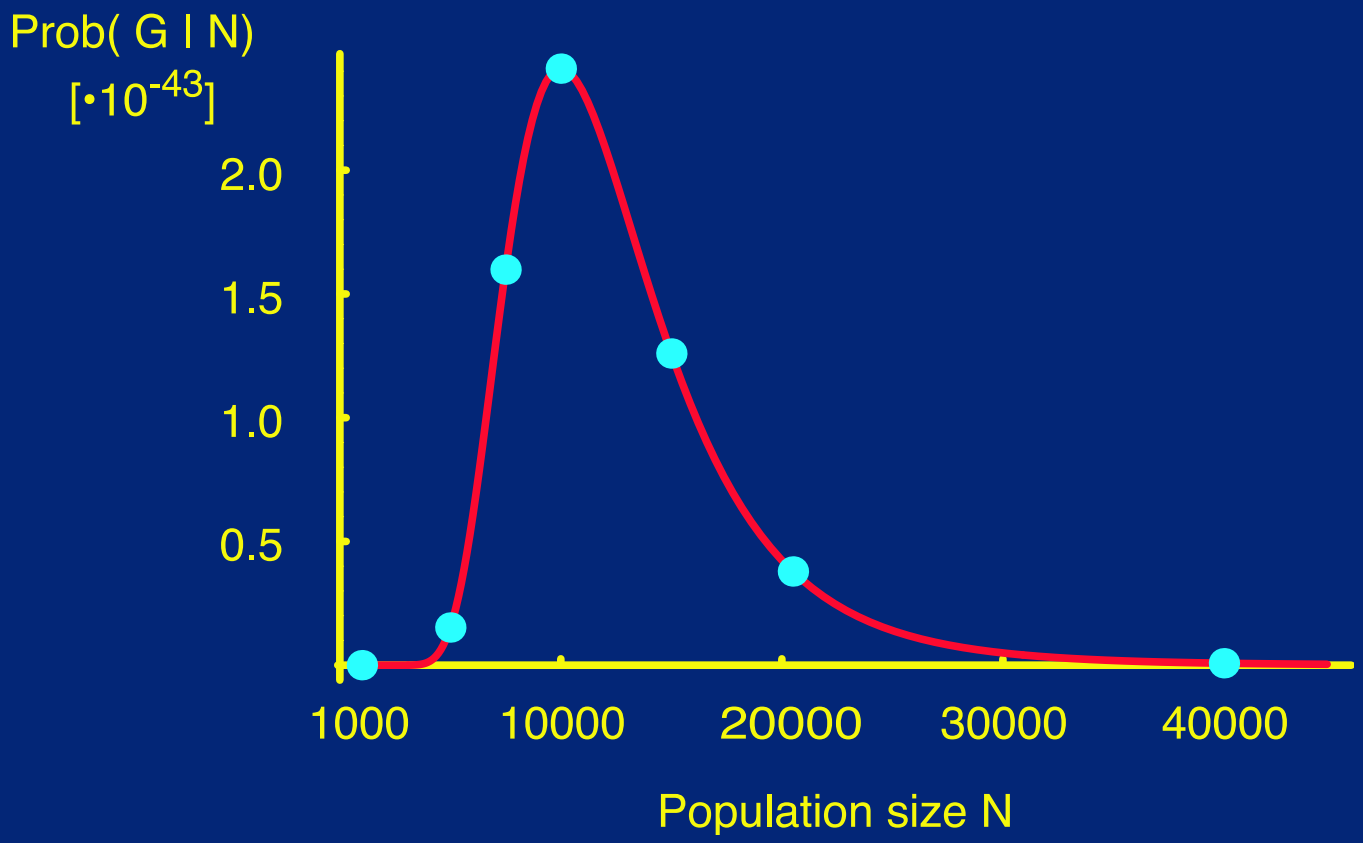
1. We get the correct genealogy from an infallible oracle

2. We remember the probability calculation

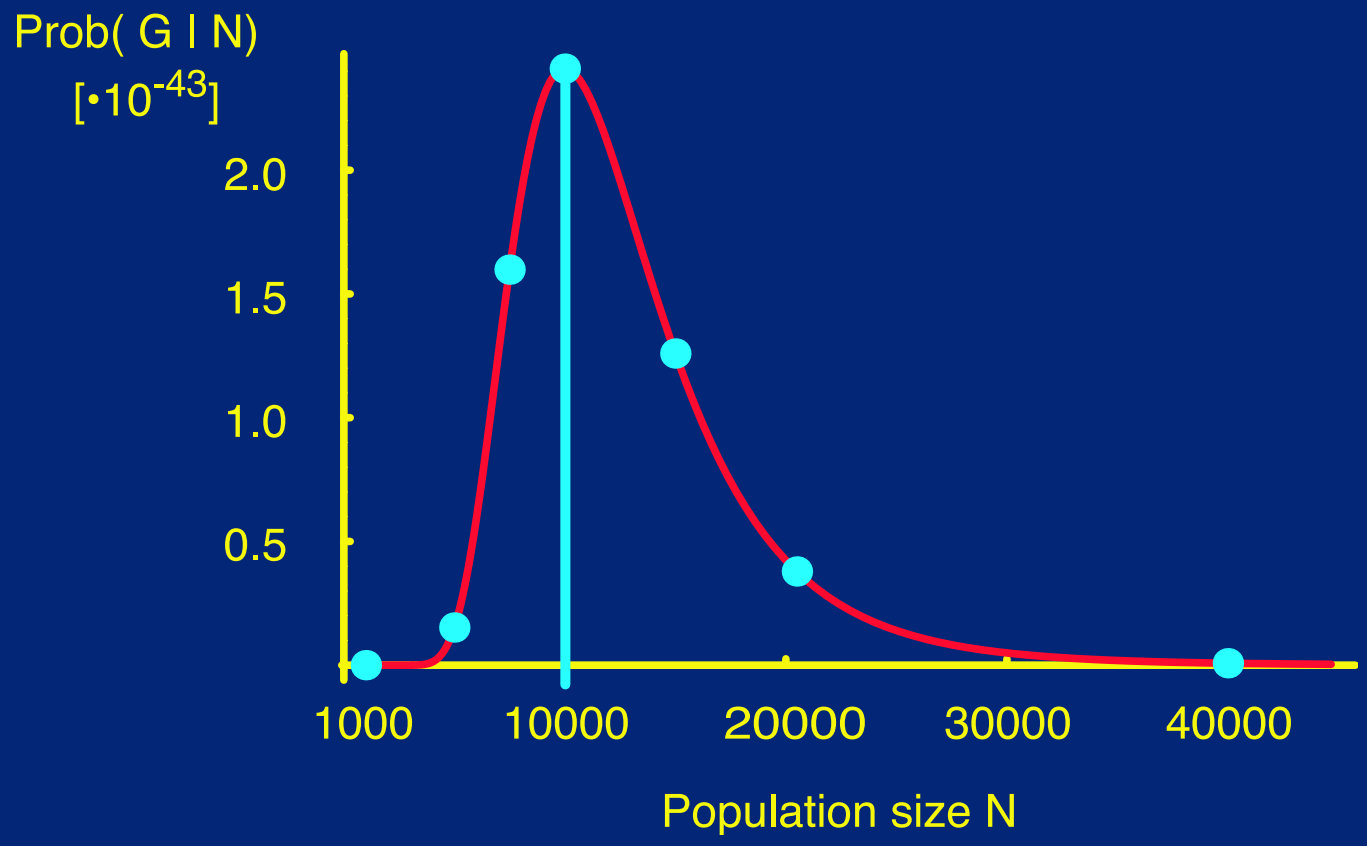


$$\mathrm{p}(\text{Genealogy}|\mathrm{N}) = \prod_{\mathrm{j}}^{\mathrm{T}} \mathrm{e}^{-\mathrm{u_j}\frac{\mathrm{k_j}(\mathrm{k_j}-1)}{4\mathrm{N}}} \frac{1}{2\mathrm{N}}$$

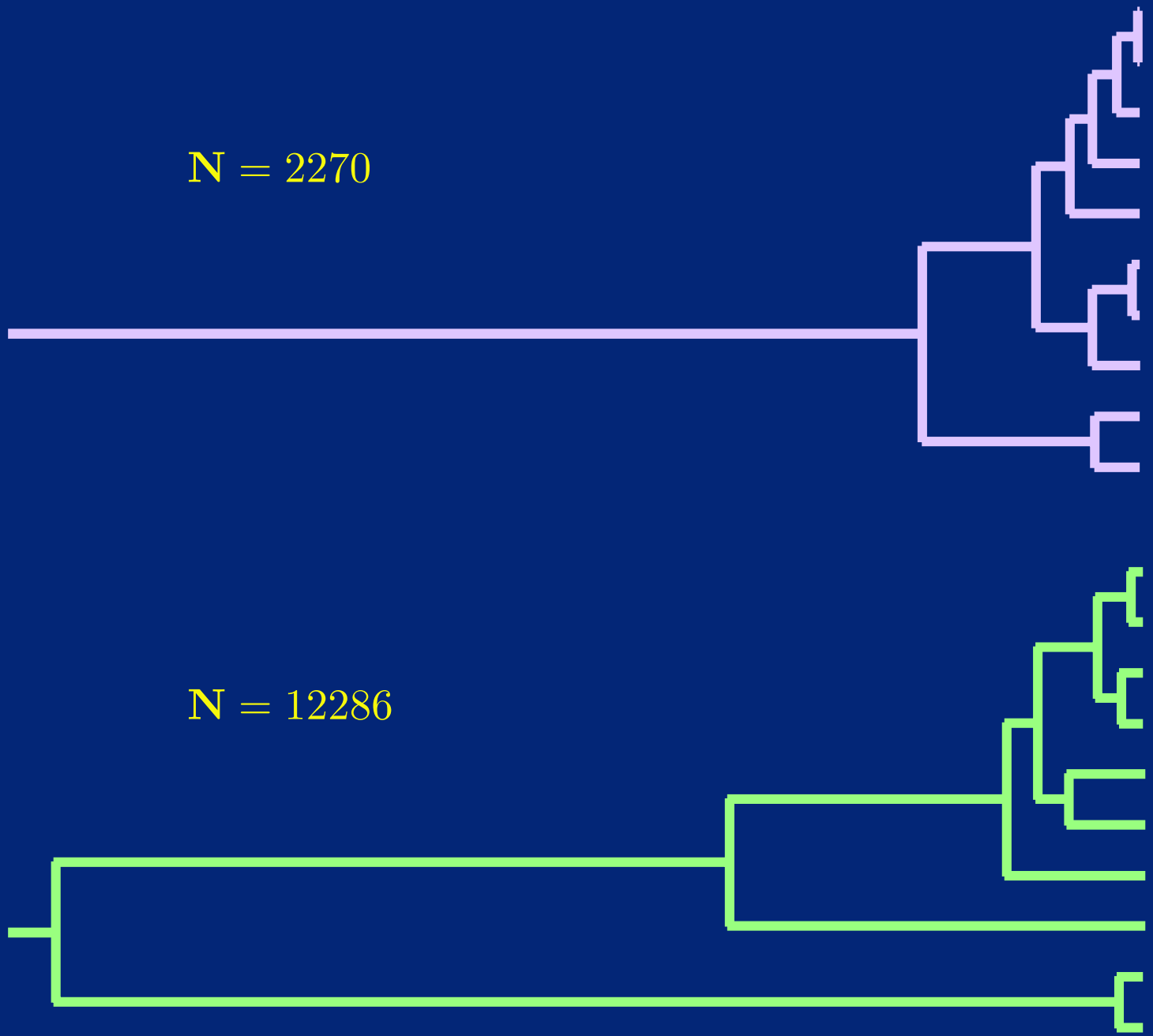
Utopian population size estimator



Utopian population size estimator



Utopian population size estimator



Lack of infallible oracles

- We assume we know the true genealogy including branch lengths
- We don't really know that
- We probably can't even infer it:
  - Tree inference is hard in general
  - Population data usually don't have enough information for good tree inference

Non-likelihood use of coalescent

- Summary statistics
  - Watterson's estimator of θ
  - FST (estimates θ and/or migration rate)
  - Hudson's and Wakeley's estimators of recombination rate
- Known-tree methods
  - UPBLUE (Yang)
  - Skyline plots (Strimmer, Pybus, Rambaut)

These methods are conceptually easy, but not always powerful, and they are difficult to extend to complex cases.

Genealogy samplers

- Acknowledge that there is an underlying genealogy–
  - but we don't know it
  - we can't infer it with high certainty
  - we can't sum over all possibilities
- A directed sample of plausible genealogies–
  - can capture much of the information in the unknown true genealogy
  - takes a long time but not forever
- These are **genealogy sampler** methods

Outline

1. Introduction to coalescent theory
2. **Practical example: red drum**
3. Genealogy samplers
4. Break
5. Survey of samplers
6. Evolutionary forces
7. Practical considerations

What is the effective population size of red drum?

Red drum, *Sciaenops ocellatus*, are large fish found in the Gulf of Mexico.

Turner, Wares, and Gold
Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish
Genetics 162:1329-1339 (2002)

What is the effective population size of red drum?

- Census population size: 3,400,000
- Effective population size: ?
- Data set:
  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population

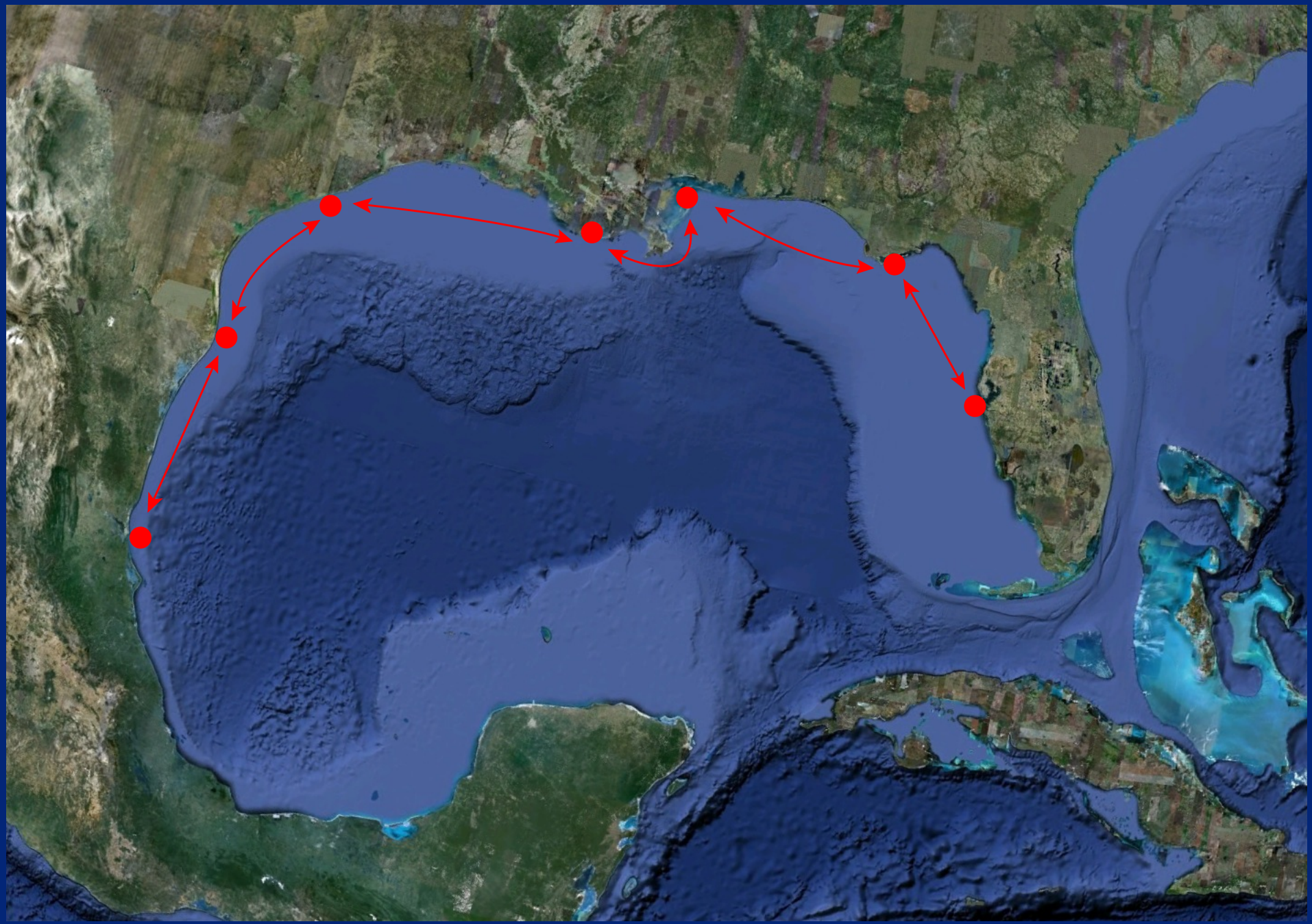
What is the effective population size of red drum?

Three approaches:

1. Allele frequency fluctuation from year to year
   - Measures current population size
   - May be sensitive to short-term fluctuations
2. Coalescent estimate from *Migrate*
   - Measures long-term harmonic mean of population size
   - May reflect past bottlenecks or other long-term effects
3. Demographic models
   - Attempt to infer genetic size from census size
   - Vulnerable to errors in demographic model
   - Not well established for long-lived species with high reproductive variability

Population model used for Migrate

- Multiple populations along Gulf coast
- Migration allowed only between adjacent populations
- Allowing for population structure should improve estimates of population size



What is the effective population size of red drum?

Estimates:

Census size (N):	3,400,000
Allele frequency method (N_e):	3,516 (1,785-18,148)
Coalescent method (N_e):	1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.

What is the effective population size of red drum?

- Allele frequency estimators measure current size
- Coalescent estimators measure long-term size
- Conclusion: population size and structure have been stable

What is the effective population size of red drum?

- Effective population size at least 1000 times smaller than census
- This result was highly surprising
- Red drum has the genetic liabilities of a rare species
- Turner et al. hypothesize an "estuary lottery"
- Unless the eggs are in exactly the right place, they all die

Outline

1. Introduction to coalescent theory
2. Practical example
3. **Genealogy samplers**
4. Break
5. Survey of samplers
6. Evolutionary forces
7. Practical considerations

Coalescent estimation of population parameters

- Mutation model: Steal a likelihood model from phylogeny inference
- Population genetics model: the Coalescent

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta)$$

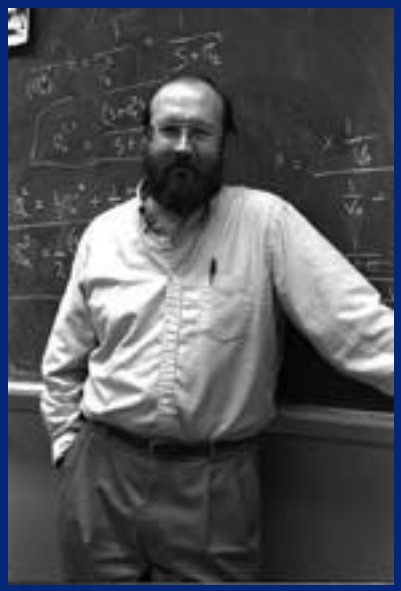
Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$P(Data|G)$ comes from a mutational model



Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_{G} P(Data|G)P(G|\Theta)$$

$P(G|\Theta)$ comes from the coalescent

Coalescent estimation of population parameters

$$L(\Theta) = P(Data|\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

$\sum_G$ is a problem

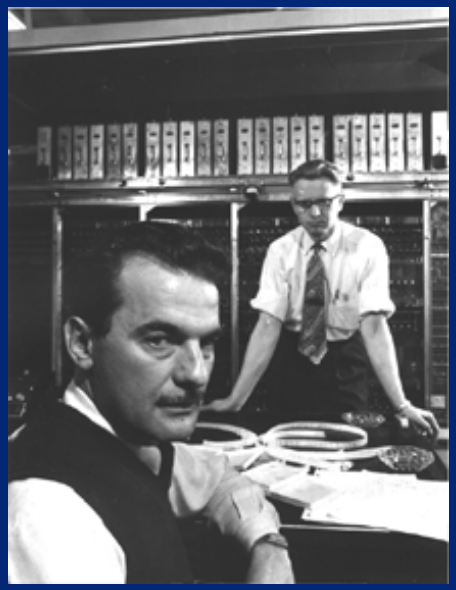
Can we calculate this sum over all genealogies?

Tips	Topologies
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	436846661310306951246468019862076389144064000000000000000
40	3027333829948007356546303364551457200042939432053862501707888721920000000000000000000
50	3.28632×10^{112}
100	1.37416×10^{284}

A solution: Markov chain Monte Carlo

- If we can't sample all genealogies, could we try a random sample?
  - Not really.
- How about a sample which focuses on good ones?
  - What is a good genealogy?
  - How can we find them in such a big search space?

A solution: Markov chain Monte Carlo



Metropolis recipe

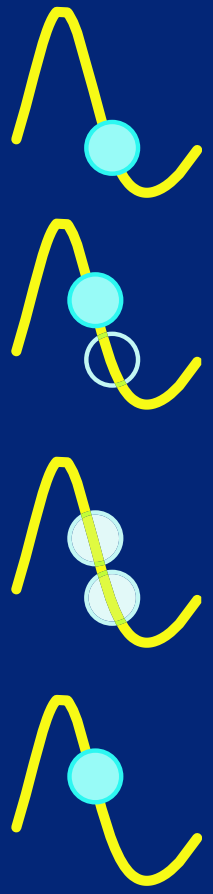
0. first state

1. perturb old state and calculate probability of new state

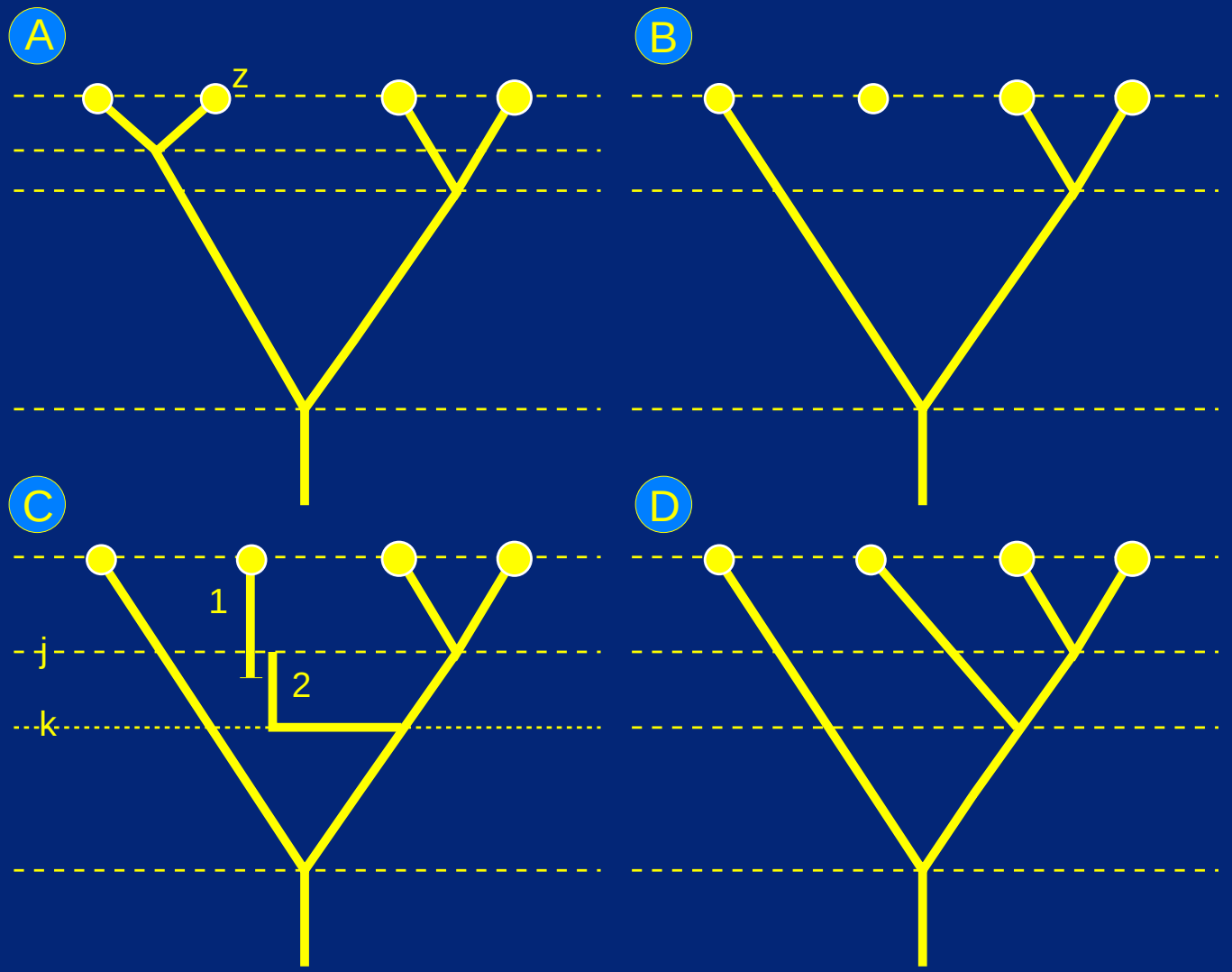
2. test if new state is better than old state: accept if ratio of new and old is larger than a random number between 0 and 1.

3. move to new state if accepted otherwise stay at old state

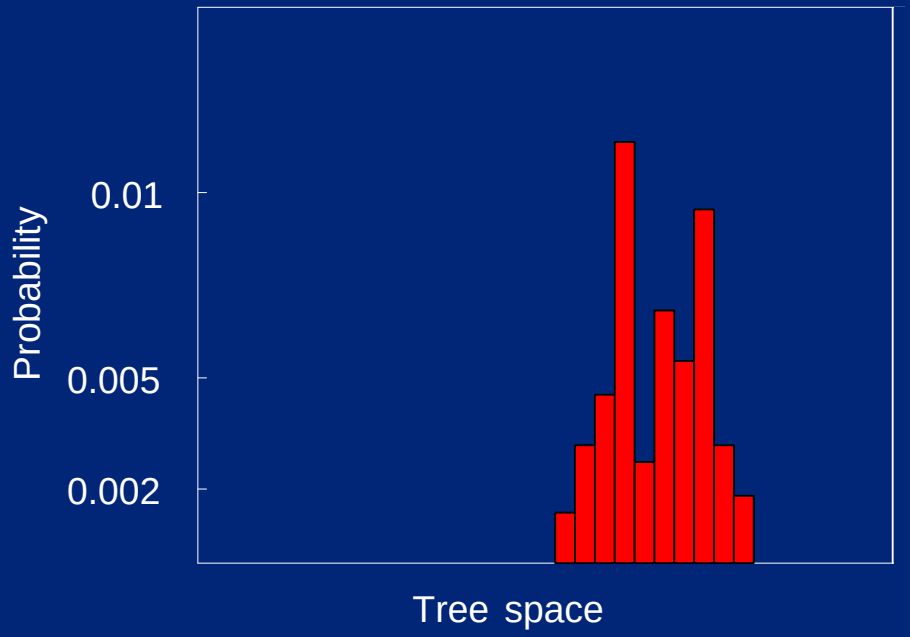
4. go to 1



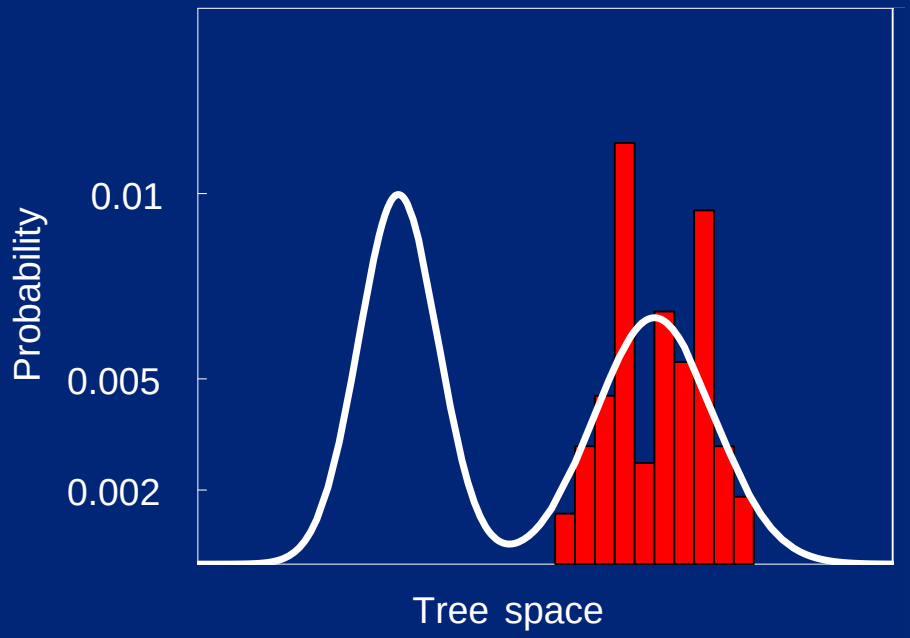
How do we change a genealogy?



MCMC walk result



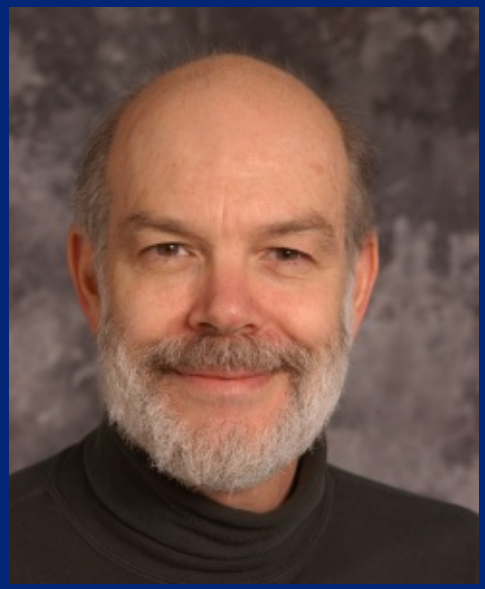
MCMC walk result–with problems



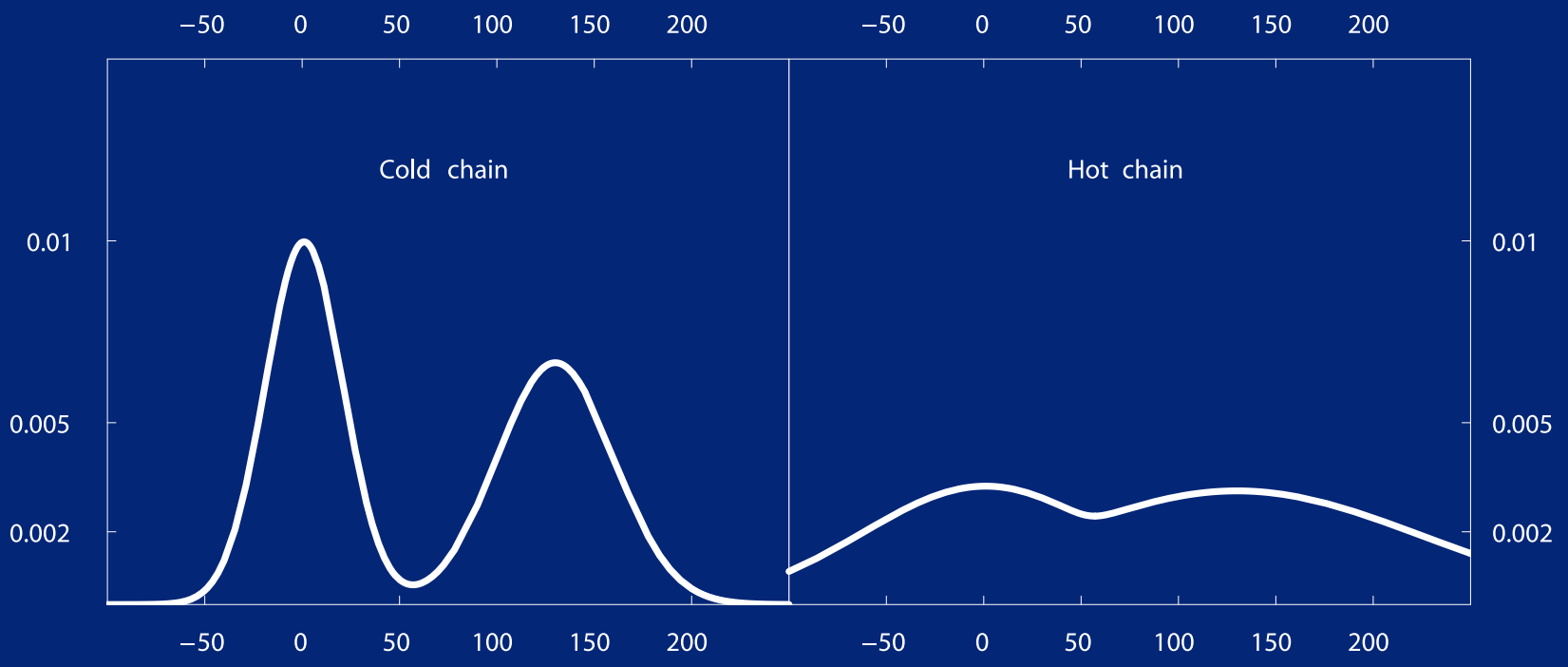
Improving our MCMC walker: Heating

Metropolis Coupled Markov chain Monte Carlo (AKA MC^3)

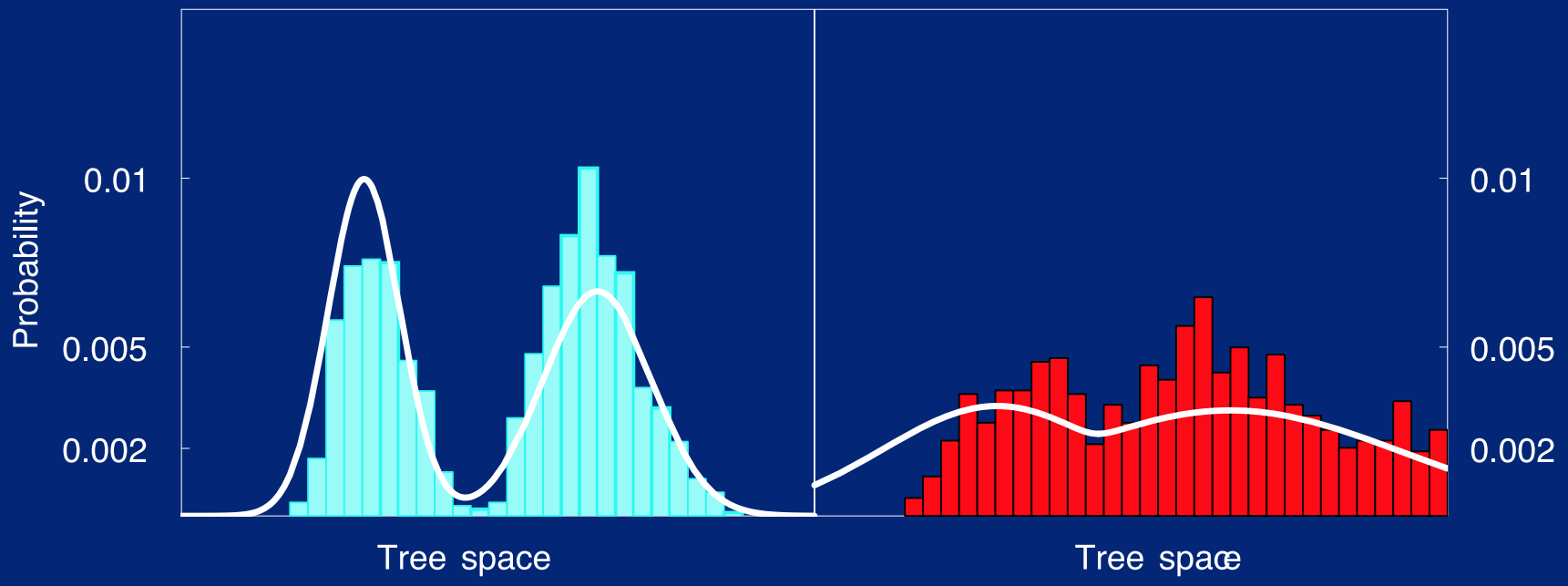
- Run several independent parallel chains: each has a different temperature
- After some sampling of genealogies, swap the genealogies of a pair of chains if the ratio between probabilities in the cold and the hot chain is larger than a random number drawn between 0 and 1.



Improving our MCMC walker: MCMCMC or MC^3



better MCMC walk result



Outline

1. Introduction to coalescent theory
2. Genealogy samplers
   (a) **Likelihood version**
   (b) **Bayesian version**
3. Practical example
4. Break
5. Survey of samplers
6. Evolutionary forces
7. Practical considerations

Likelihood and Bayesian approaches

- All genealogy samplers search among genealogies
- All of them require some type of guide value ("driving value") to determine which genealogies will be proposed
- Two major approaches: Likelihood-based and Bayesian
- Major ideological difference, relatively small practical one

Likelihood samplers

- Use arbitrary values of the parameters to guide the search
- Sample genealogies throughout the search
- At the end of the search, evaluate $P(G|\Theta)$ for sampled genealogies
- Correct for the influence of the driving values
- Iterate to improve driving values

Bayesian samplers

- Propose new driving values throughout the run
- New driving values drawn from a prior
- Accept or reject driving values based on $P(G|\Theta)$
- Final conclusions based on histogram of driving values

Likelihood analysis

We will approximate:

$$L(\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

Likelihood analysis

We will approximate:

$$L(\Theta) = \sum_G P(Data|G)P(G|\Theta)$$

by sampling n genealogies from $P(Data|G)P(G|\Theta_0)$:

$$L(\Theta) = \frac{1}{n}\sum_{G^*} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Here the G^* are no longer random genealogies; they are sampled from a distribution that depends on the **driving value** Θ_0

Likelihood analysis

$$L(\Theta) = \frac{1}{n} \sum_{G} \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Isn't this circular? We have a solution for the unknown $L(\Theta)$ in terms of the unknown $L(\Theta_0)$.

Likelihood analysis

$$L(\Theta) = \frac{1}{n} \sum_G \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)/L(\Theta_0)}$$

Isn't this circular? We have a solution for the unknown $L(\Theta)$ in terms of the unknown $L(\Theta_0)$.

$$\frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n} \sum_G \frac{P(Data|G)P(G|\Theta)}{P(Data|G)P(G|\Theta_0)}$$

This doesn't give us the actual value of $L(\Theta)$ but it does allow us to compare various values of Θ and choose the best.

Likelihood analysis

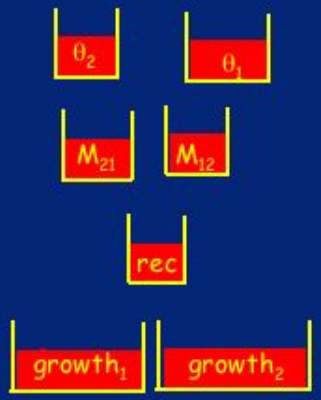
- This approach is only asymptotically correct
- For finite sample sizes, it has a bias toward its driving value
- We can greatly reduce this:
  - Start with an arbitrary Θ_0
  - Run the sampler a while and estimate the best Θ
  - It will be biased toward Θ_0, but...
  - Use it as the new Θ_0 and start over

Bayesian approach

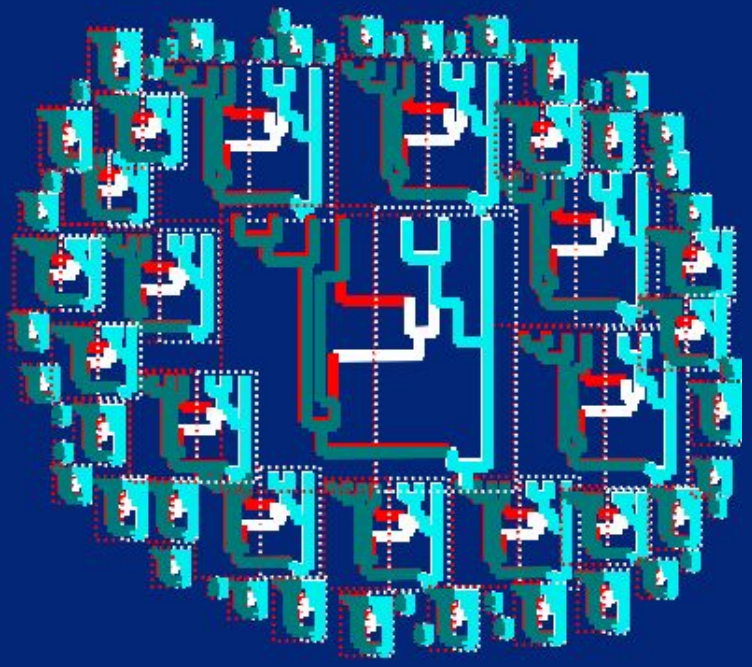
- A Bayesian analysis requires us to provide priors for all parameters
- These $could$ be based on detailed knowledge of the biology
- In practice, uninformative flat priors are used

New search scheme for Bayes

Parameter space
(determined by priors)



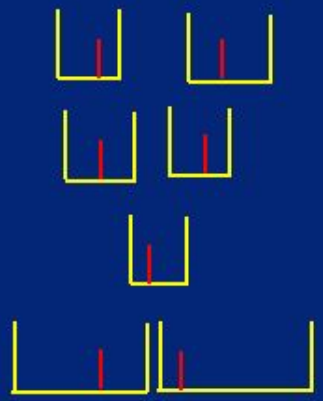
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

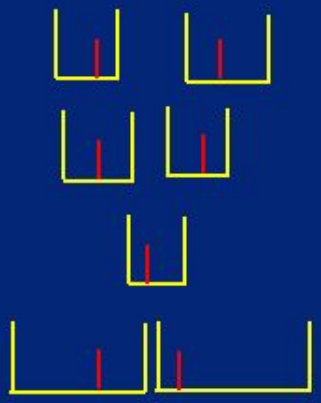
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

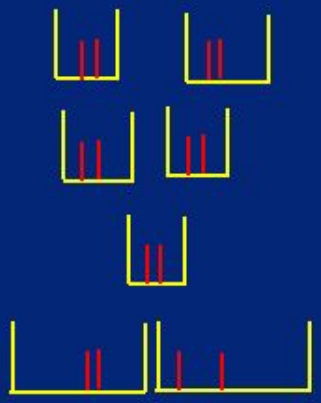
Tree space



New search scheme for Bayes

Parameter space
(determined by priors)

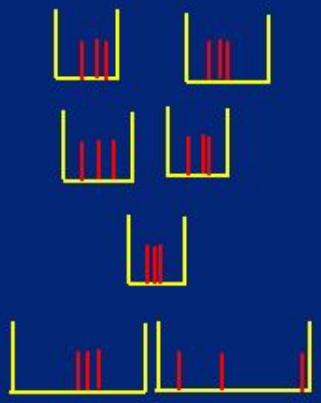
Tree space



New search scheme for Bayes

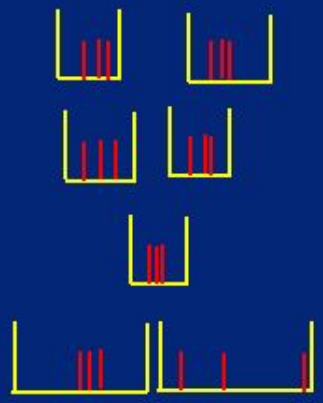
Parameter space
(determined by priors)

Tree space



New search scheme for Bayes

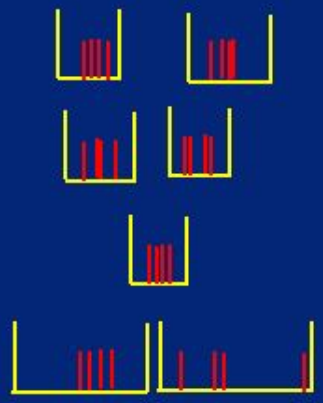
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

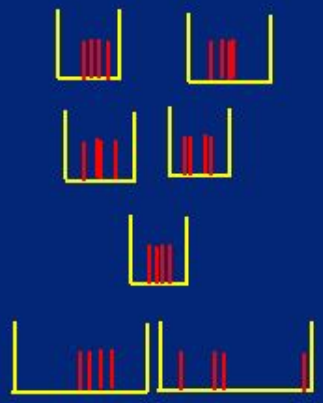
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

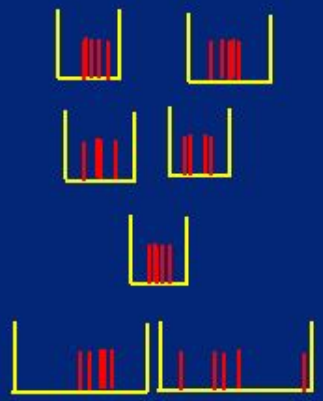
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

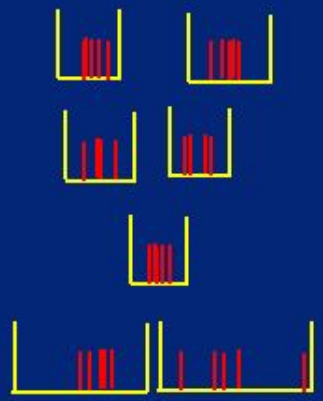
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

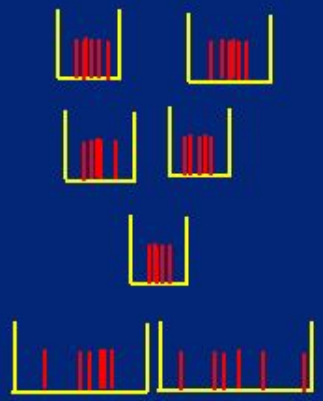
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

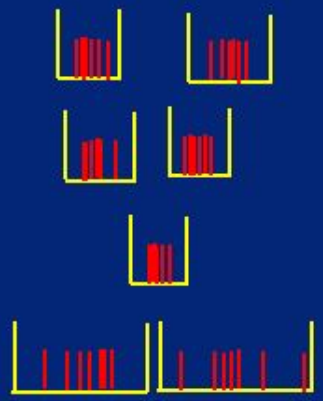
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

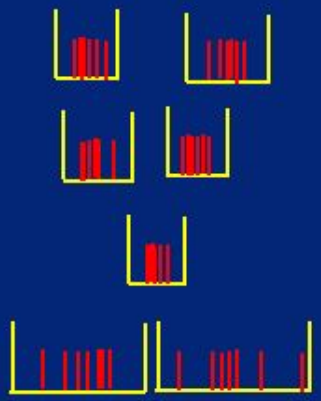
Parameter space
(determined by priors)



Tree space

New search scheme for Bayes

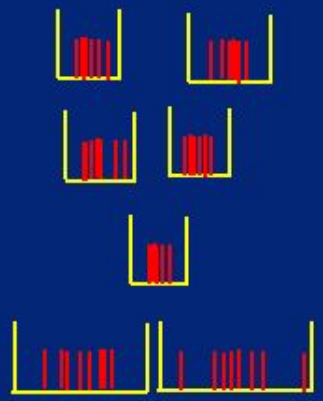
Parameter space
(determined by priors)



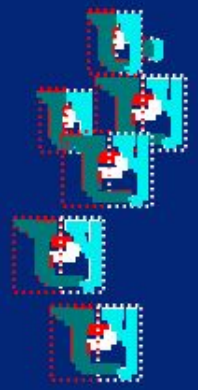
Tree space

New search scheme for Bayes

Parameter space
(determined by priors)

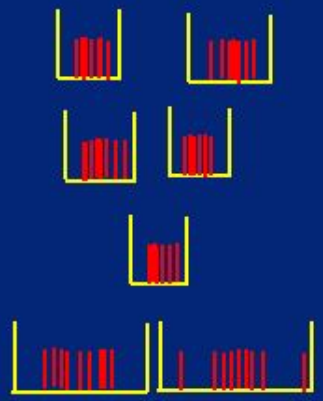


Tree space

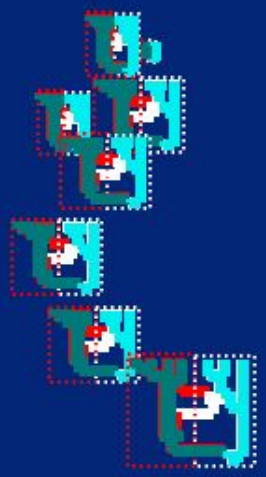


New search scheme for Bayes

Parameter space
(determined by priors)

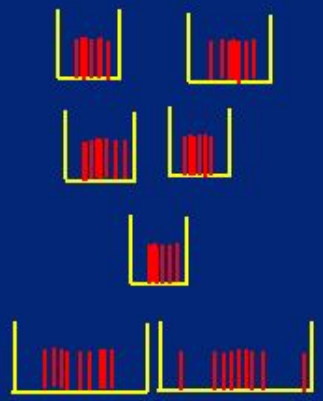


Tree space

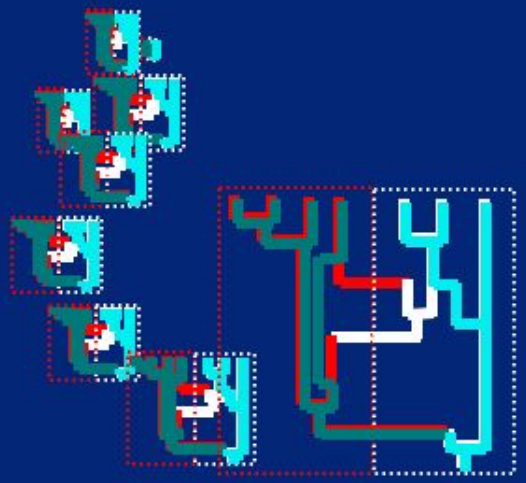


New search scheme for Bayes

Parameter space
(determined by priors)

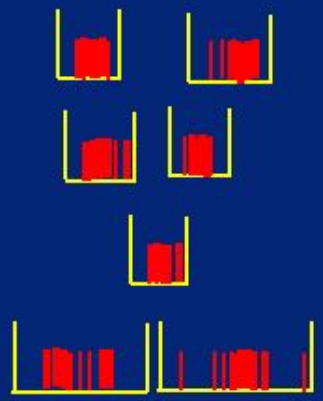


Tree space

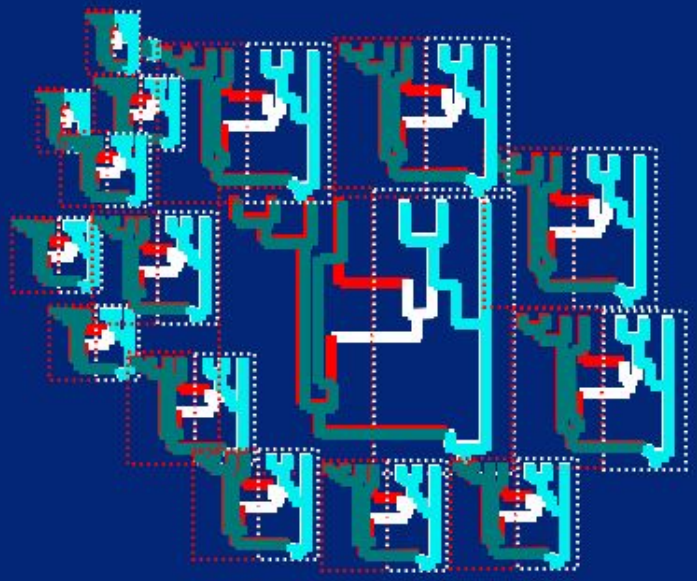


New search scheme for Bayes

Parameter space
(determined by priors)

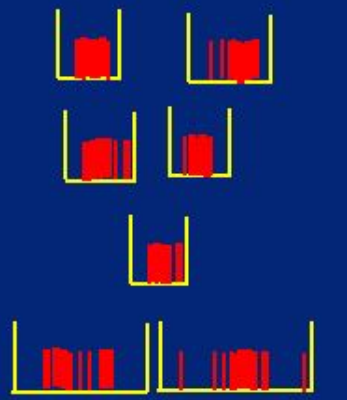


Tree space

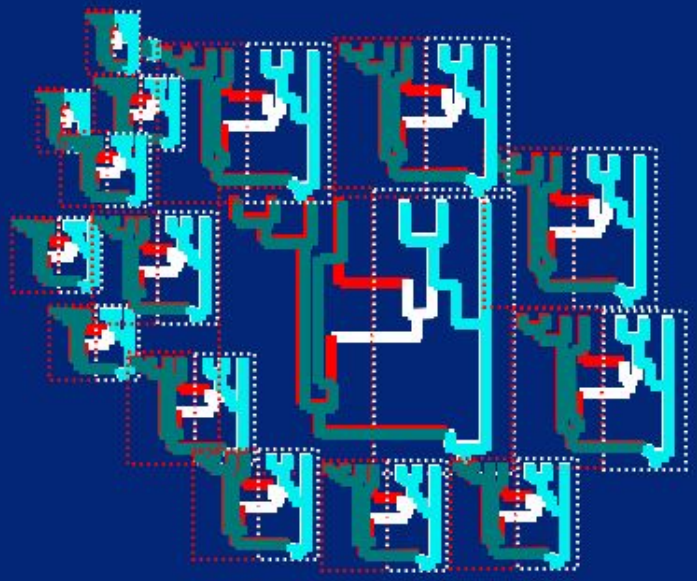


New search scheme for Bayes

Parameter space
(determined by priors)



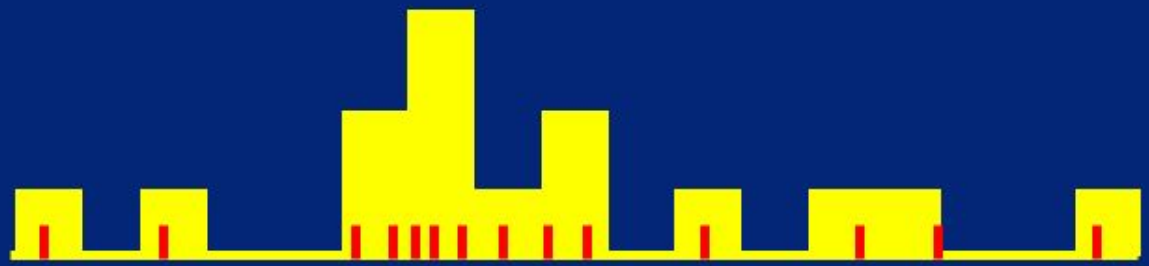
Tree space



Keep a list of all accepted parameters

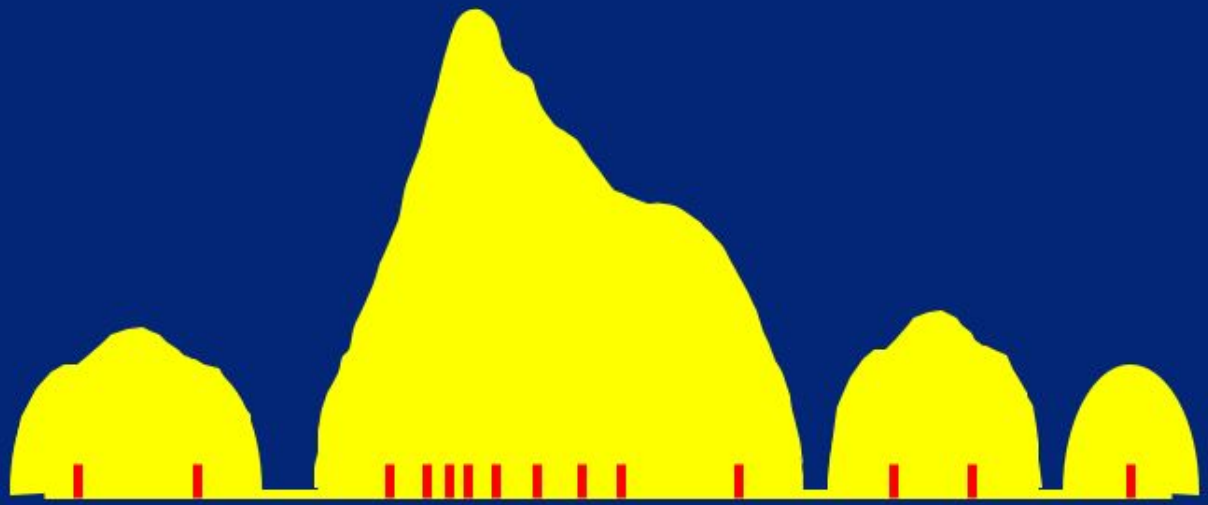
Data collection and curve smoothing

Data collection and curve smoothing

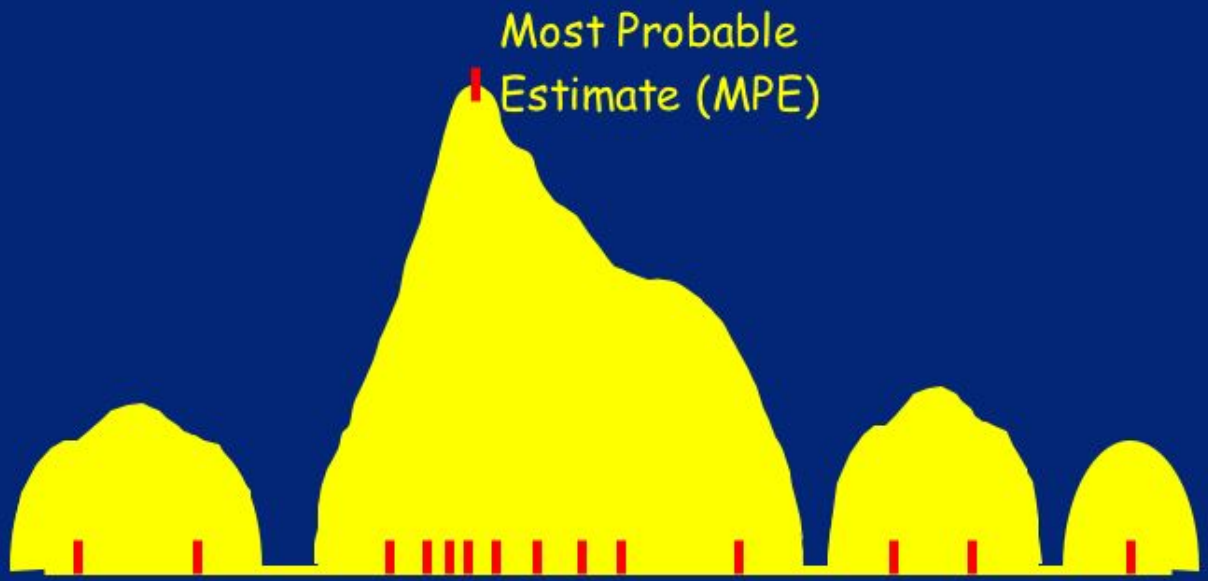


Data collection and curve smoothing

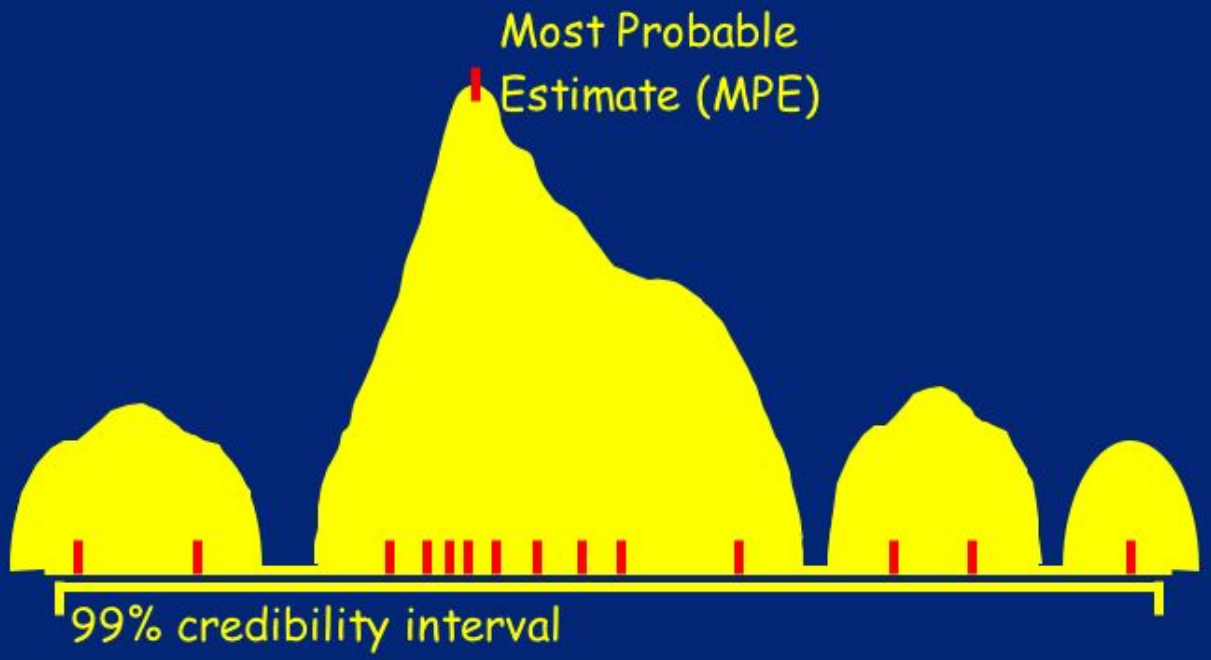
Data collection and curve smoothing



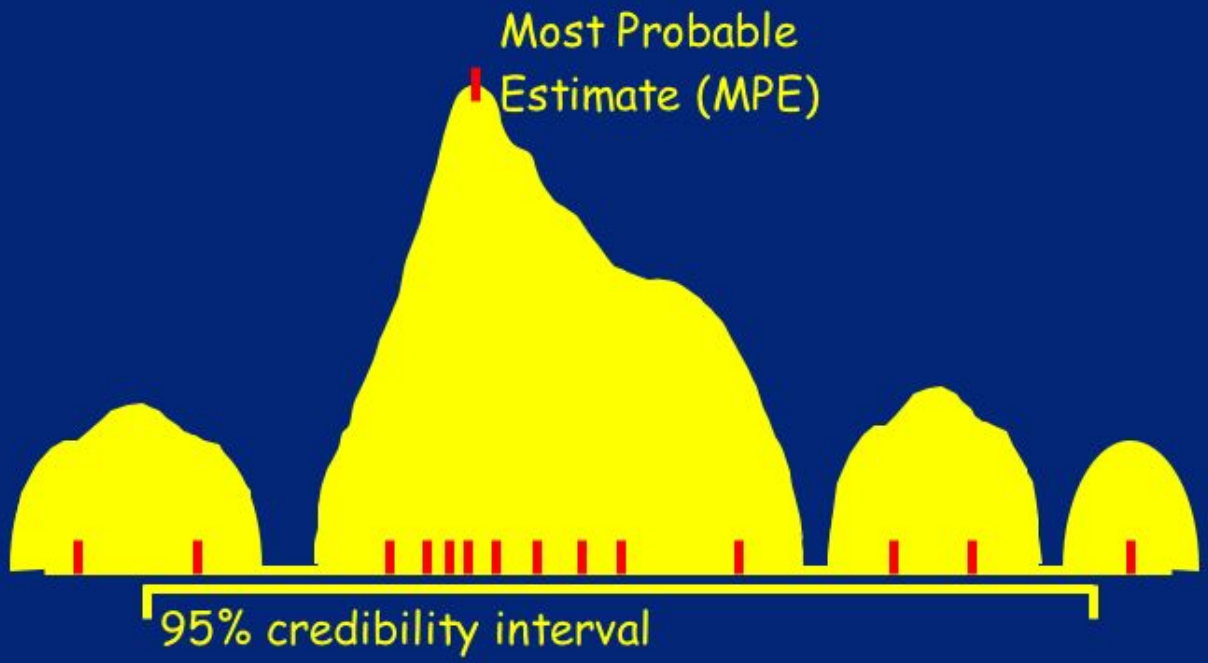
Data collection and curve smoothing



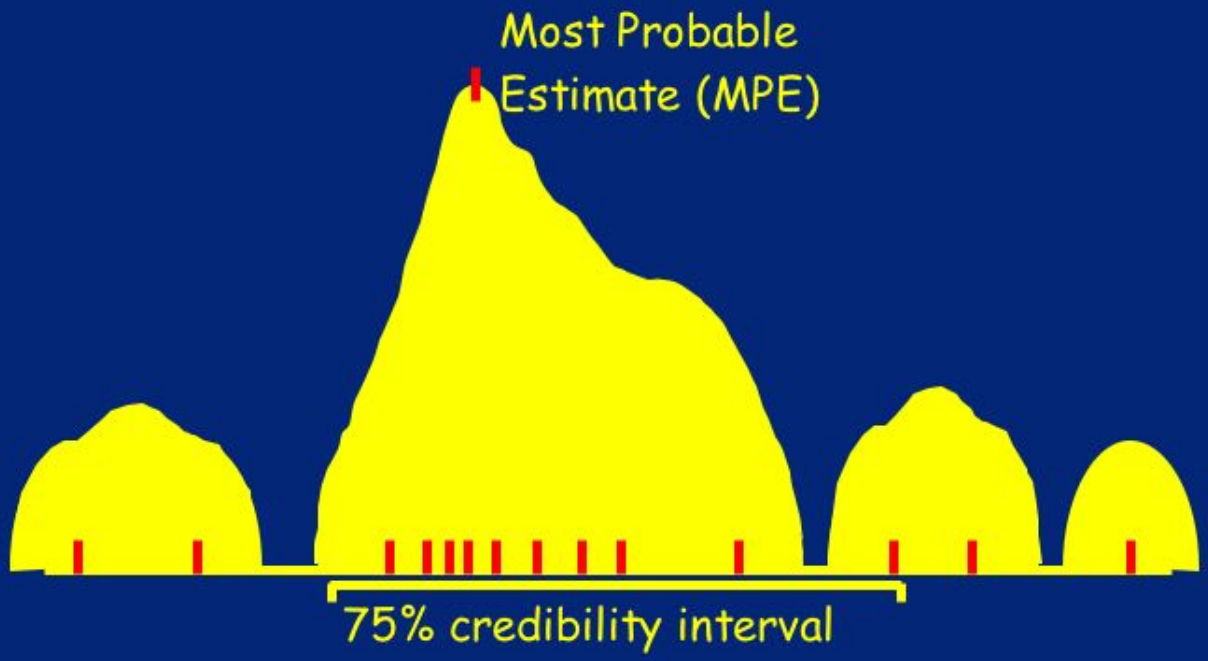
Data collection and curve smoothing



Data collection and curve smoothing



Data collection and curve smoothing



Advantages of Bayesian analysis

- Easier to interpret probabilities than likelihoods
- Smoothing a histogram is quicker than finding maxima of a likelihood curve
- Not dependent on starting driving values
- Parameter values near zero estimated more accurately
- Prior information can be incorporated (in theory)
- Trendy!

Disdvantages of Bayesian analysis

- No information currently available on correlation of parameters
- Dependent on good priors; results can be severely distorted by bad priors

Bottom line

- Kuhner 2006: Bayes and likelihood almost identical
- Beerli 2006: Bayes has edge with sparse data
- My recommendations:
  - Use Bayes if you think a parameter is very close to zero
  - Otherwise, with rich data either method is good
  - With poor data, do you really want to be doing this analysis at all?
  - When using Bayes, be careful of your priors!
- If the genealogy search is inadequate, both methods will fail (and fail in similar ways)

Break