

Thomas Mailund

Bioinformatics Research Center

University of Aarhus

*Joint work with Asger Hobolth, Ole F. Christiansen and
Mikkel H. Schierup*

Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model

Asger Hobolth^{1*}, Ole F. Christensen², Thomas Mailund^{2,3}, Mikkel H. Schierup²

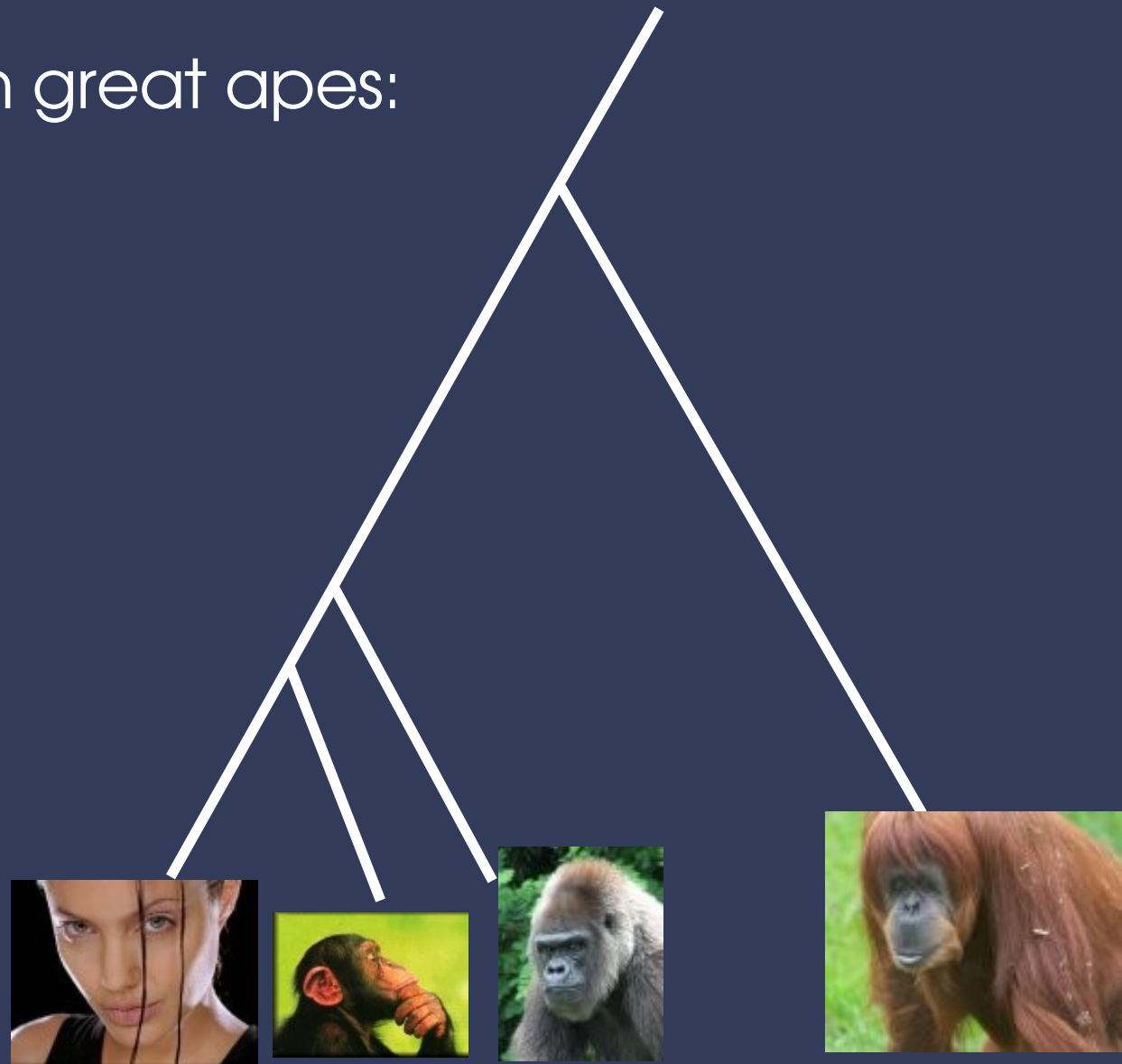
1 Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Bioinformatics Research Center, University of Aarhus, Aarhus, Denmark, **3** Department of Statistics, University of Oxford, Oxford, United Kingdom

The genealogical relationship of human, chimpanzee, and gorilla varies along the genome. We develop a hidden Markov model (HMM) that incorporates this variation and relate the model parameters to population genetics quantities such as speciation times and ancestral population sizes. Our HMM is an analytically tractable approximation to the coalescent process with recombination, and in simulations we see no apparent bias in the HMM estimates. We apply the HMM to four autosomal contiguous human–chimp–gorilla–orangutan alignments comprising a total of 1.9 million base pairs. We find a very recent speciation time of human–chimp (4.1 ± 0.4 million years), and fairly large ancestral effective population sizes ($65,000 \pm 30,000$ for the human–chimp ancestor and $45,000 \pm 10,000$ for the human–chimp–gorilla ancestor). Furthermore, around 50% of the human genome coalesces with chimpanzee after speciation with gorilla. We also consider 250,000 base pairs of X-chromosome alignments and find an effective population size much smaller than 75% of the autosomal effective population sizes. Finally, we find that the rate of transitions between different genealogies correlates well with the region-wide present-day human recombination rate, but does not correlate with the fine-scale recombination rates and recombination hot spots, suggesting that the latter are evolutionarily transient.

Citation: Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet 3(2): e7. doi:10.1371/journal.pgen.0030007

Dating speciation events

Relationship between great apes:



Dating speciation events

Relationship between great apes:

How do we know this is the relationship?



Dating speciation events

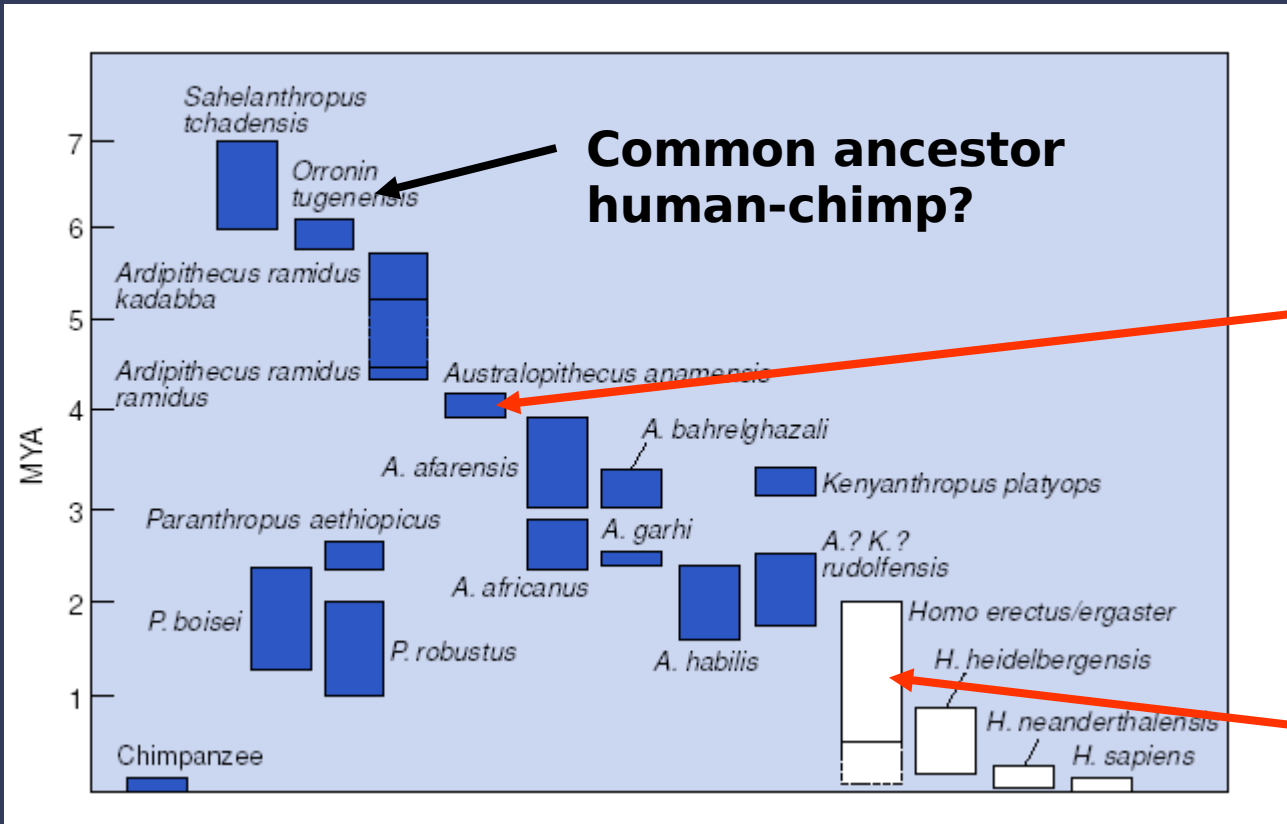
Relationship between great apes:

How do we know this is the relationship?

How do we date the speciation events?



Fossil record

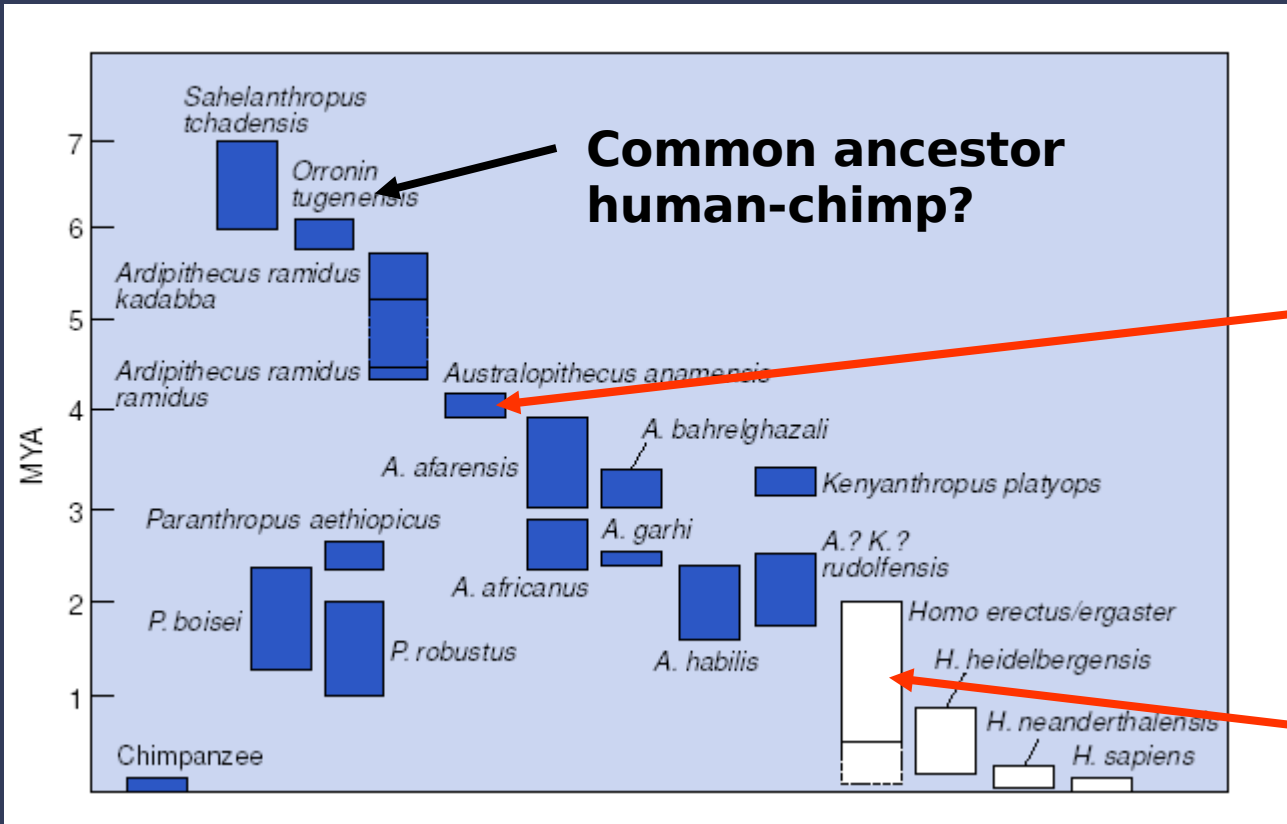


Jobling et al. 2004

Good fossil record for humans, less so for other apes...



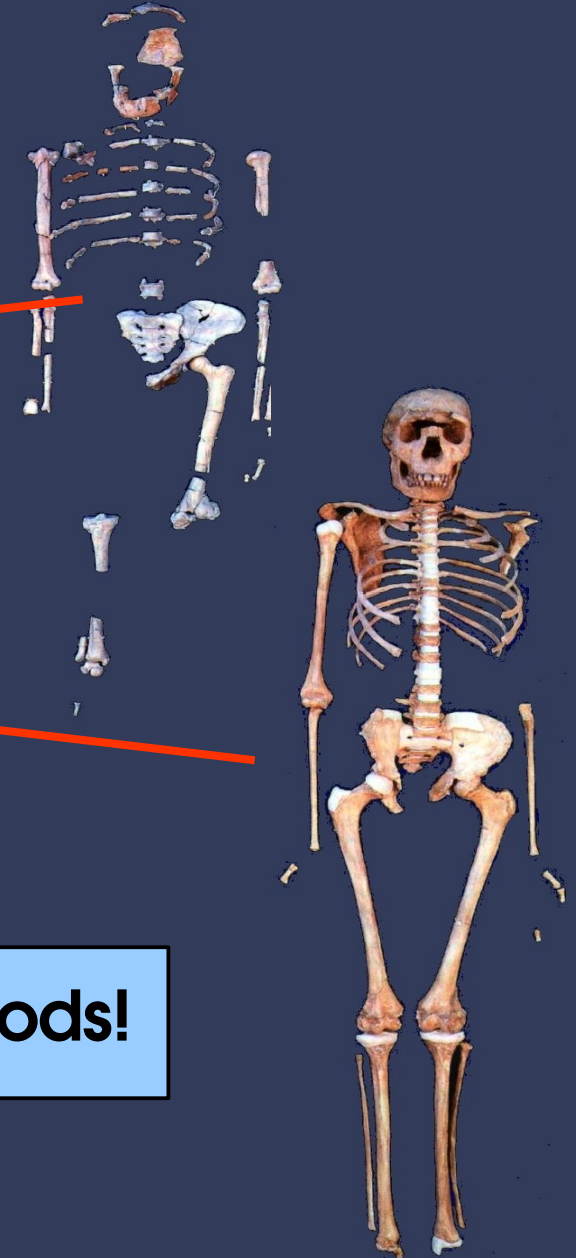
Fossil record



Jobling et al. 2004

Enter: molecular genetics methods!

Good fossil record
for other apes...



Outline of talk



- Introduction to *molecular evolution* and *population genetics*
- Mathematical modeling of the problem: *coalescence hidden Markov models*
- Inference and results

Mutations and a molecular clock



- Mutations enter DNA when
 - cells replicate
 - chromosomes are modified by the cell's "chemical soup"
- Error correction and various "proof reading" mechanisms makes mutations extremely rare
 - But there is **a lot** of DNA in each individual (2 x ~3 billion nucleotides)
 - Things that happen about every thousand years happens a lot over millions of years

Mutations and a molecular clock



Linear in time and (germ line cell) generations:

$$\mathbb{E}[m] = \mu_T \cdot t + \mu_G \cdot g$$

...and for our purposes, generations are linear in time, so:

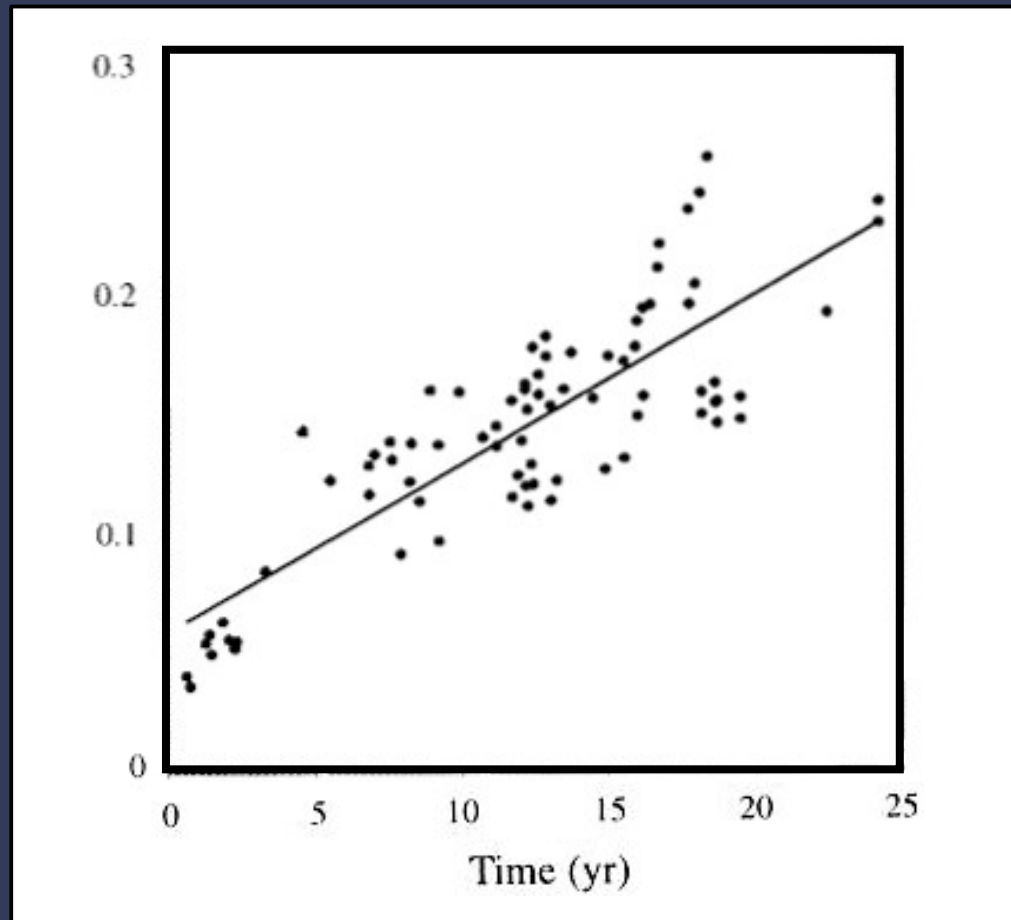
$$\mathbb{E}[m] = \mu_T \cdot t$$

Mutations and a molecular clock



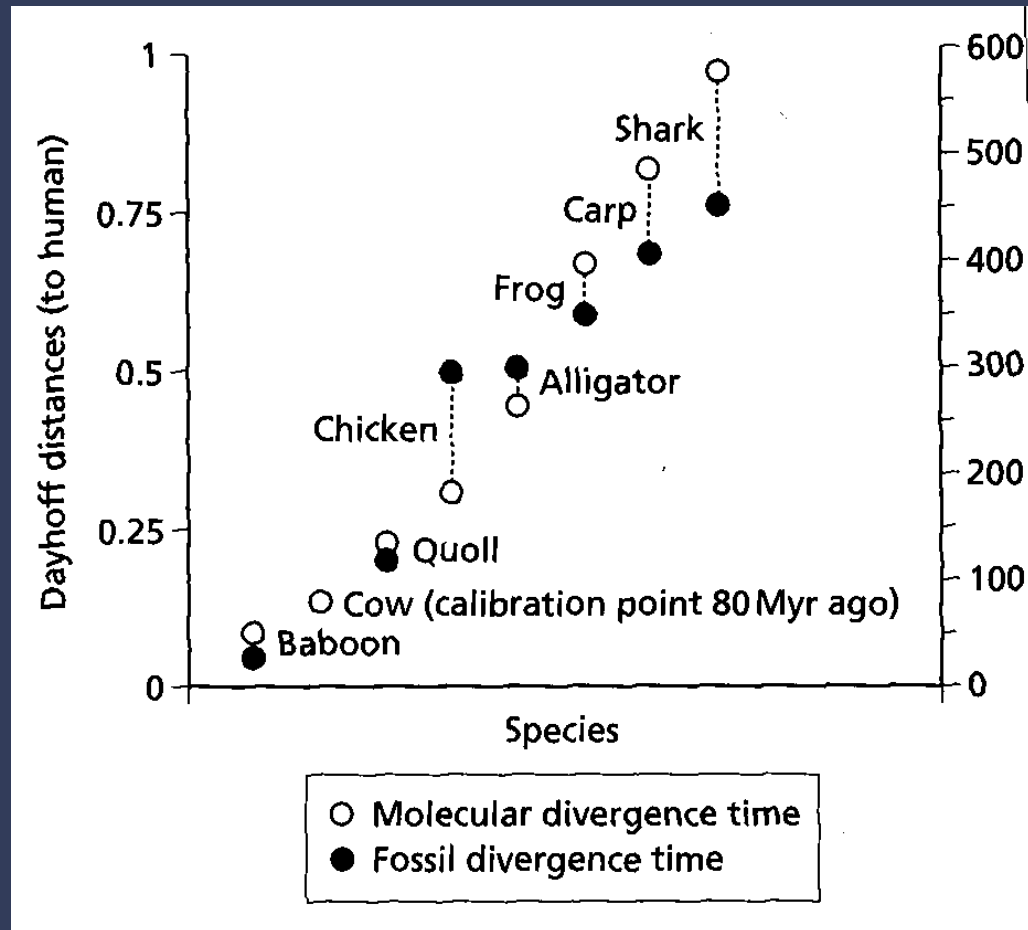
$$\mathbb{E}[m] = \mu_T \cdot t$$

HIV sequence:
Accumulation
of mutations
over time



Mutations and a molecular clock

$$\mathbb{E}[m] = \mu_T \cdot t$$



Mutations and a molecular clock



$$\mathbb{E}[m] = \mu_T \cdot t$$

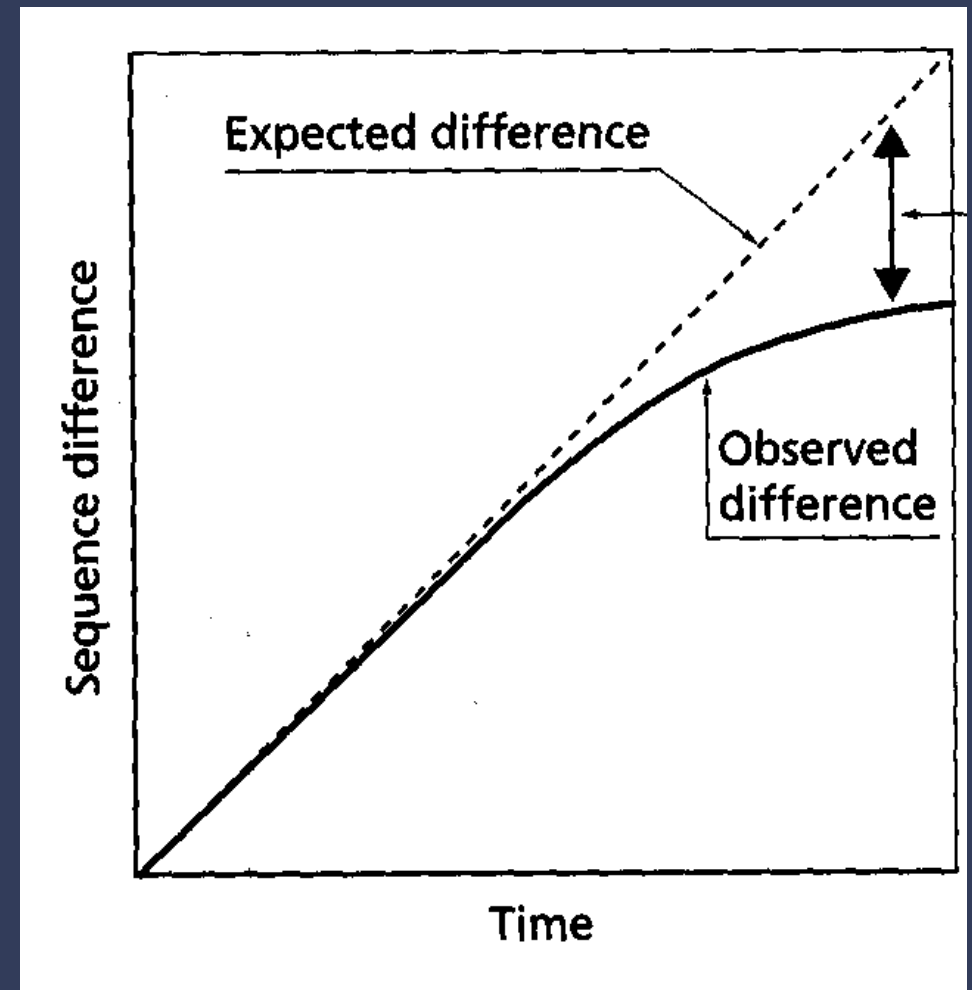
Estimation using
observed number
of mutations

$$\hat{t} = \frac{\mathbb{E}[m]}{\mu_T}$$
$$\approx \frac{m_{\text{obs}}}{\mu_T}$$

Mutations and a molecular clock



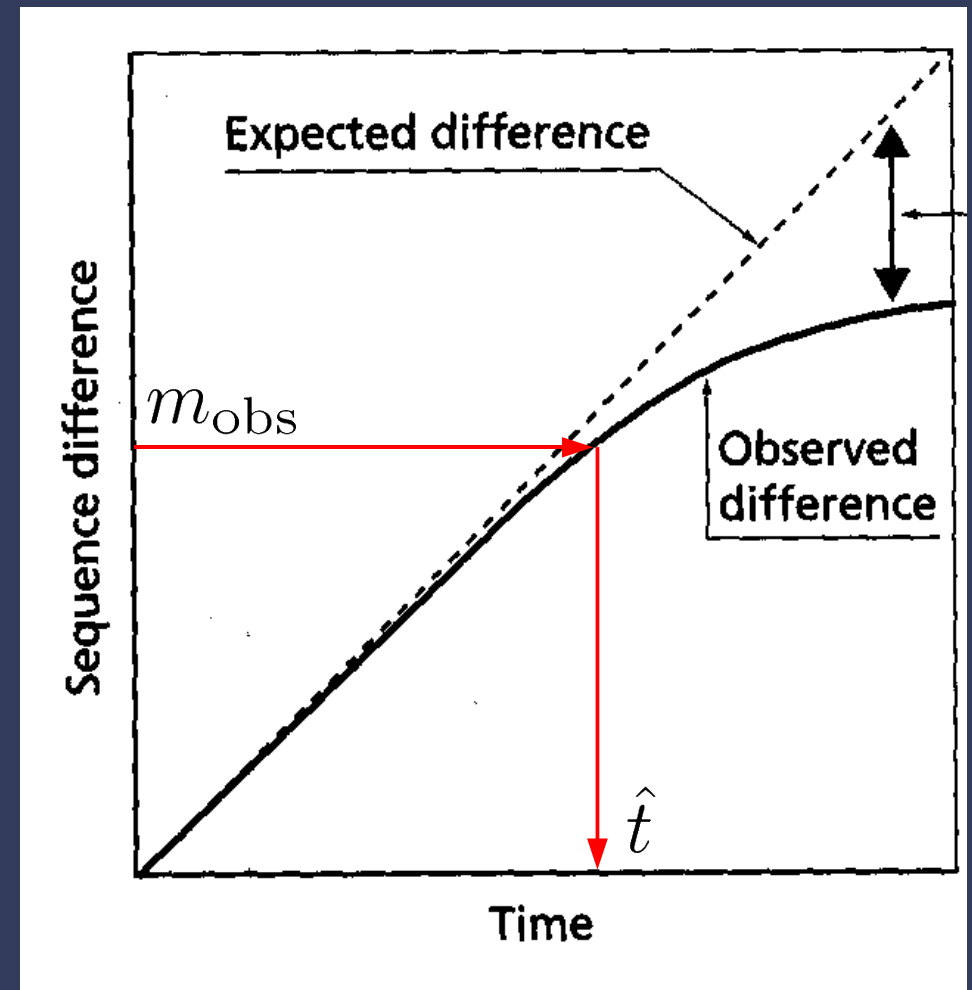
Back mutations complicate matters slightly, but we can compensate for this by solving a simple differential equation...



Mutations and a molecular clock



Back mutations complicate matters slightly, but we can compensate for this by solving a simple differential equation...



Mutations and a molecular clock



The time estimate is known up to a mutation-rate factor...

$$\hat{t} = \frac{\mathbb{E}[m]}{\mu_T}$$

...that e.g. can be calibrated using fossil evidence.

Dating divergence...

So we can estimate *pairwise* divergence in units of time...



Dating divergence...

So we can estimate *pairwise* divergence in units of time...

...although this over-estimates the number of mutations (does not take shared mutations into account)



Dating divergence...

So we can estimate *pairwise* divergence in units of time...

...possible to construct a tree and infer branch lengths from this...

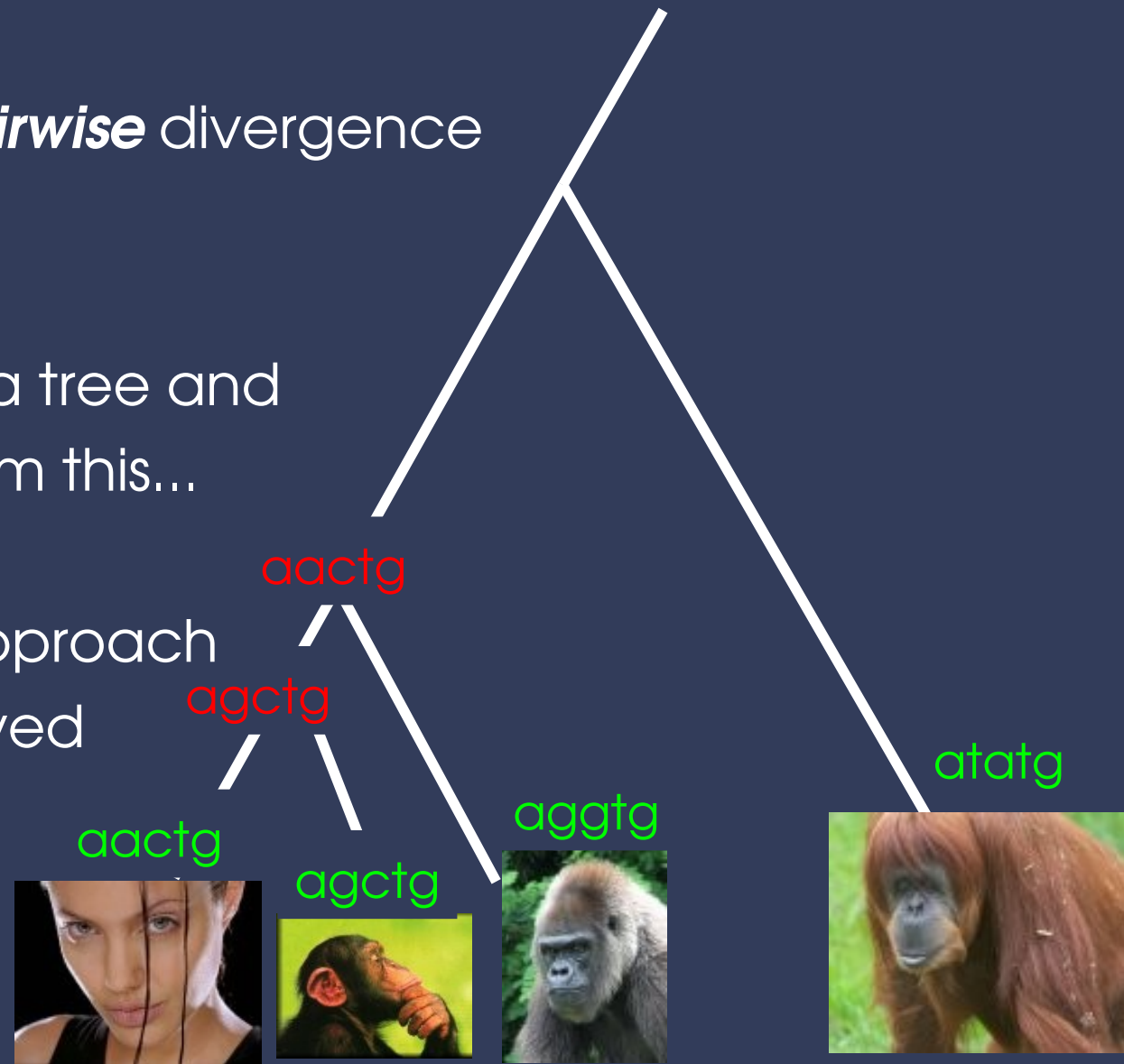


Dating divergence...

So we can estimate *pairwise* divergence in units of time...

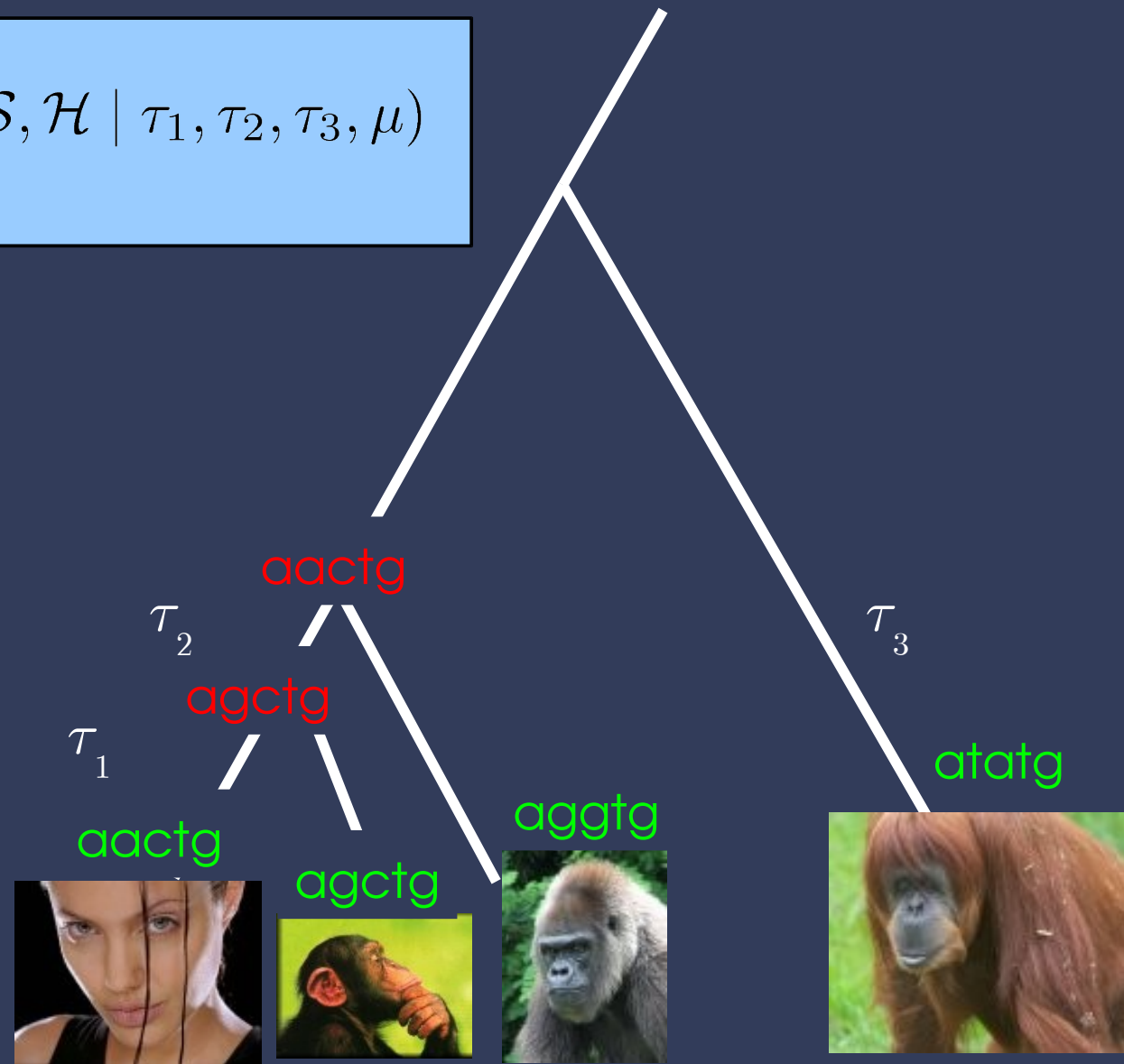
...possible to construct a tree and infer branch lengths from this...

...or take a statistical approach and deal with unobserved sequences in a mathematical model.



Dating divergence...

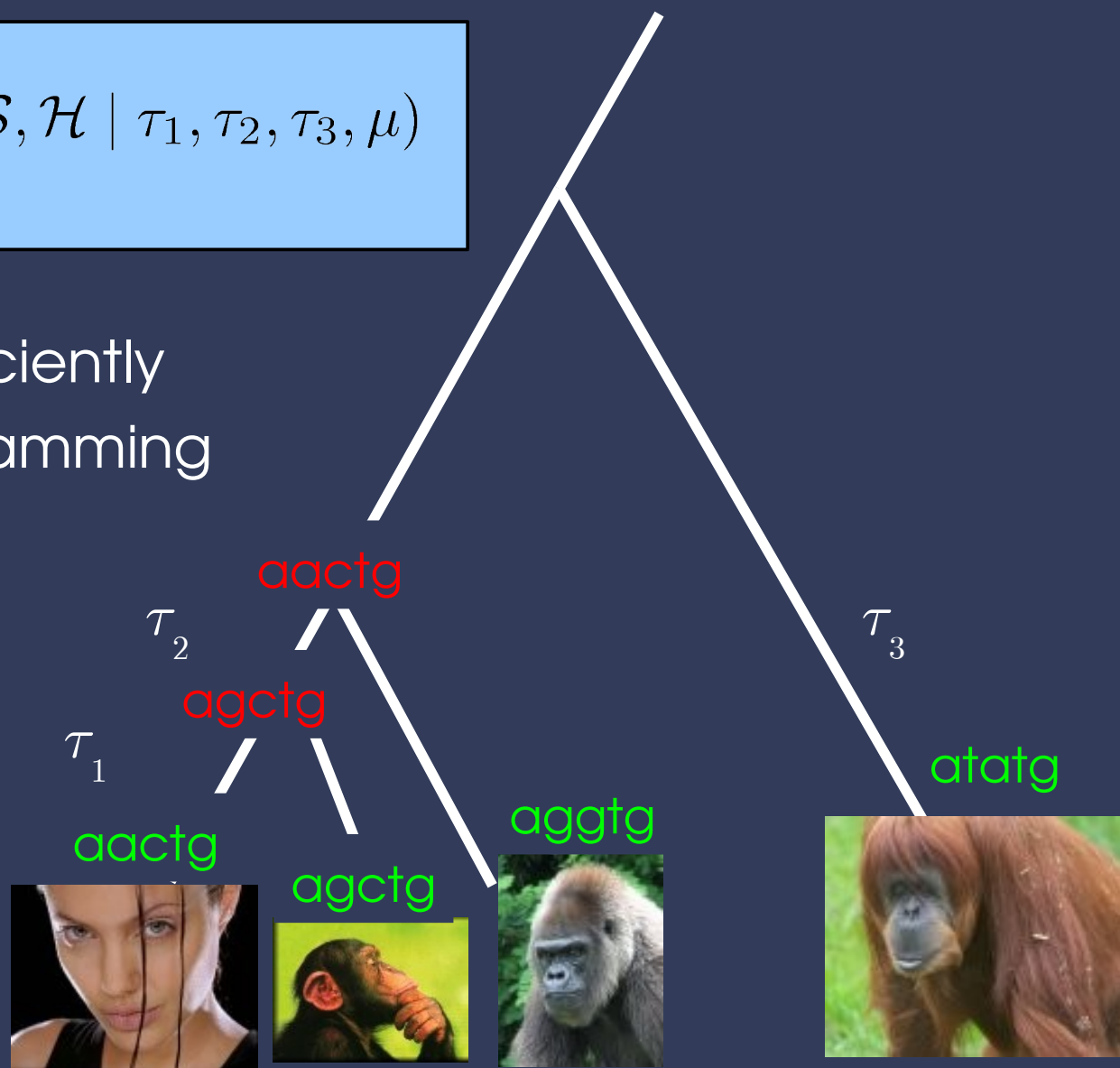
$$p(\mathcal{S} \mid \tau_1, \tau_2, \tau_3, \mu) = \sum_{\mathcal{H}} p(\mathcal{S}, \mathcal{H} \mid \tau_1, \tau_2, \tau_3, \mu)$$



Dating divergence...

$$p(\mathcal{S} \mid \tau_1, \tau_2, \tau_3, \mu) = \sum_{\mathcal{H}} p(\mathcal{S}, \mathcal{H} \mid \tau_1, \tau_2, \tau_3, \mu)$$

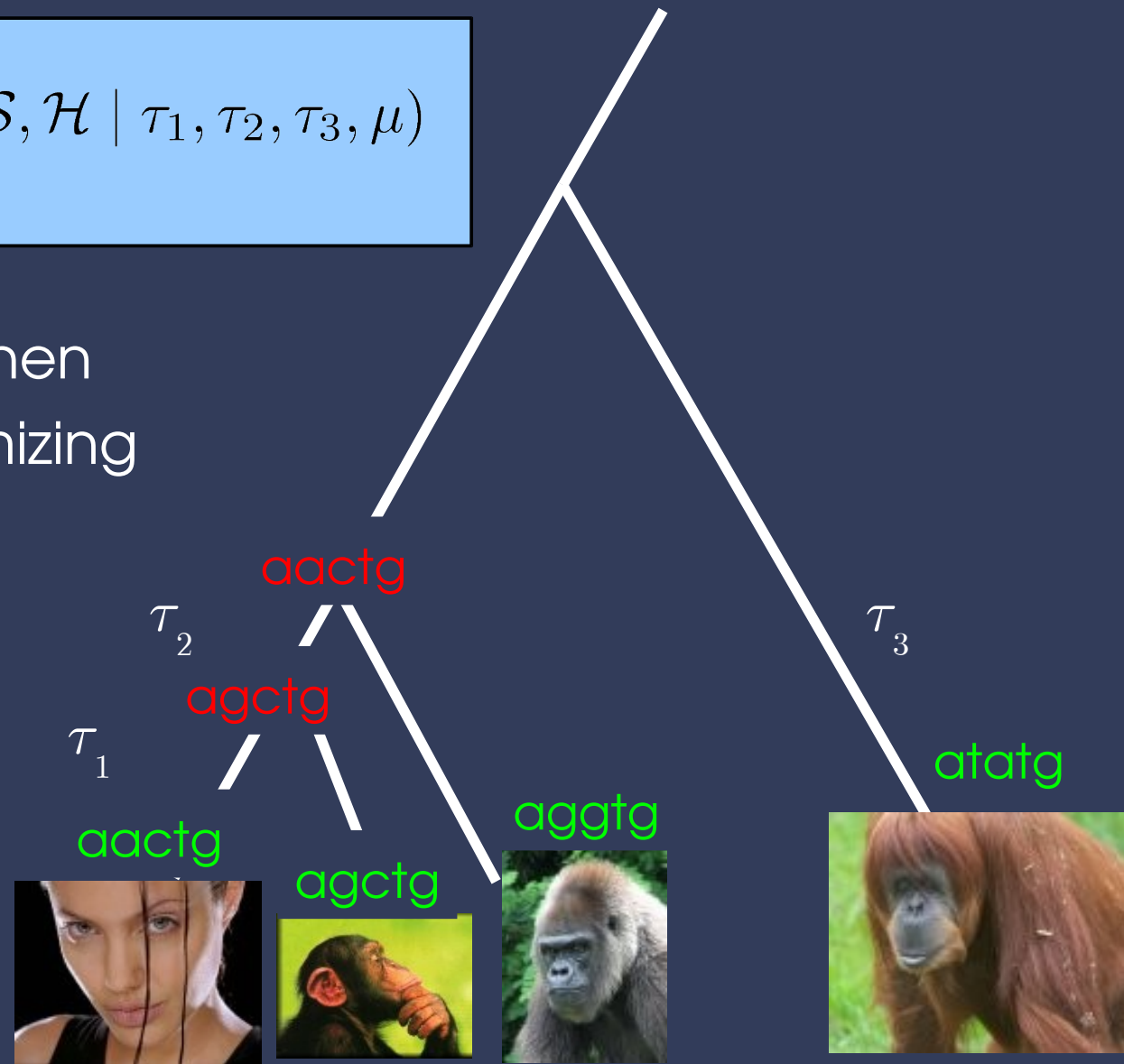
Can be computed efficiently using a dynamic programming algorithm.



Dating divergence...

$$p(\mathcal{S} \mid \tau_1, \tau_2, \tau_3, \mu) = \sum_{\mathcal{H}} p(\mathcal{S}, \mathcal{H} \mid \tau_1, \tau_2, \tau_3, \mu)$$

Time parameters can then be estimated by maximizing the likelihood.

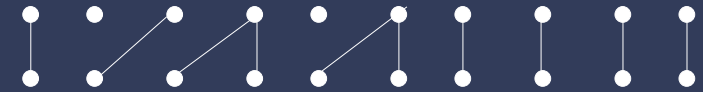


Mutations in a population



Wright-Fisher model

- Discrete, non-overlapping generations
- Constant population size
- Each individual in one generation is a random copy of an individual from the previous generation



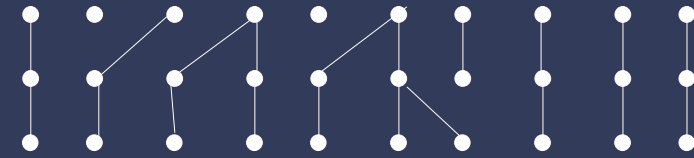
N_e individuals

Mutations in a population



Wright-Fisher model

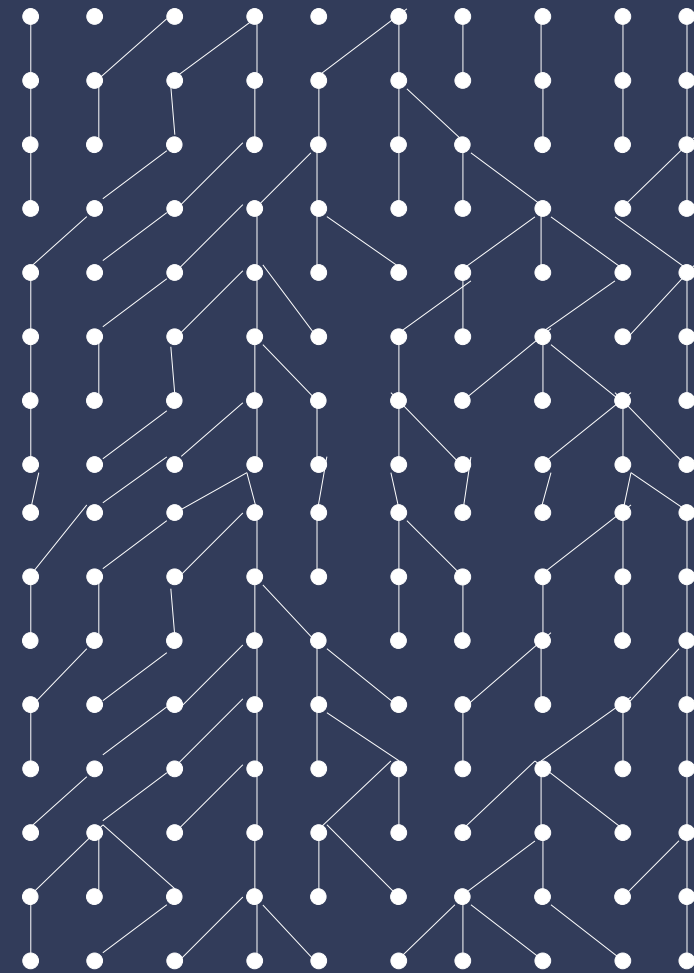
- Discrete, non-overlapping generations
- Constant population size
- Each individual in one generation is a random copy of an individual from the previous generation



N_e individuals

Wright-Fisher model

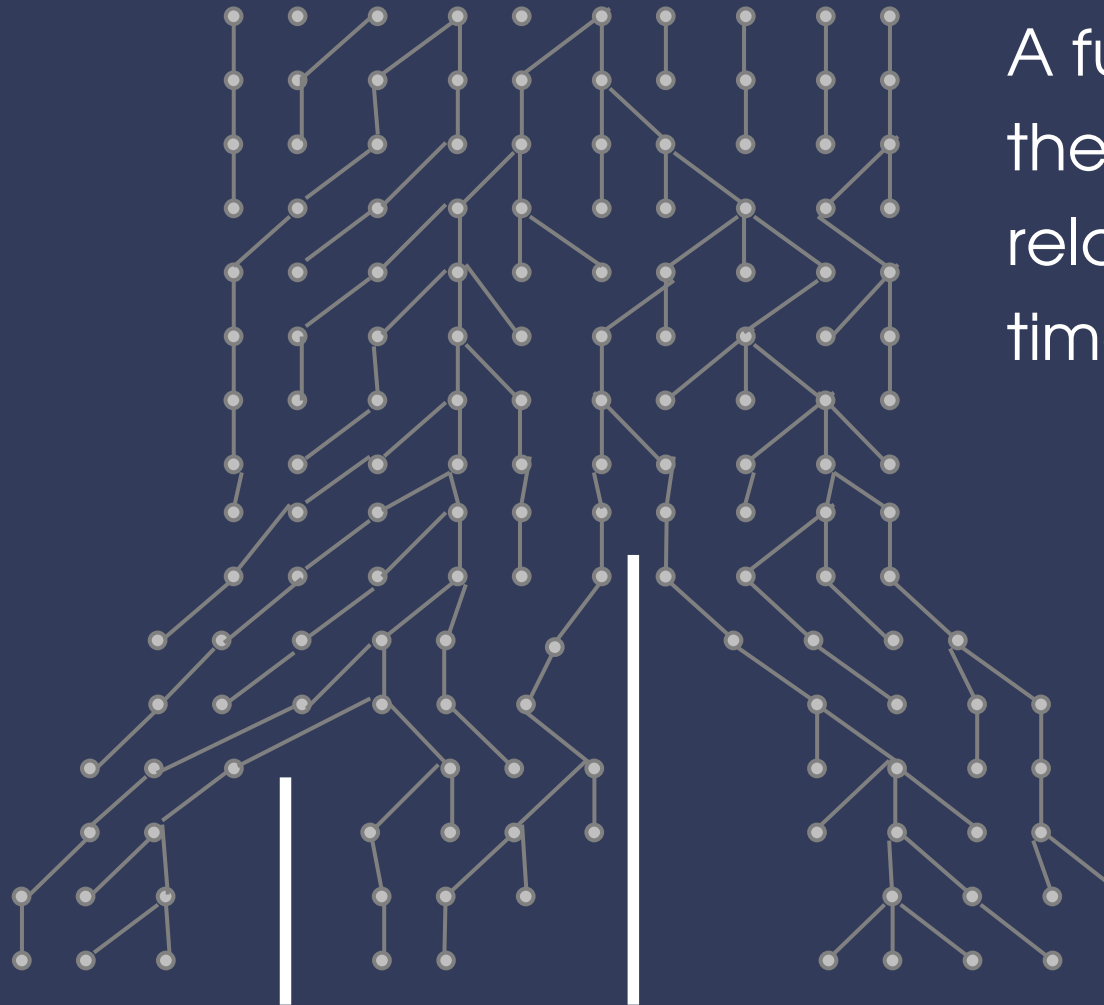
- Discrete, non-overlapping generations
- Constant population size
- Each individual in one generation is a random copy of an individual from the previous generation



N_e individuals

Populations and species

Individuals



A funny thing happens if the population size is large relative to the splitting time...

Populations and species



Individuals



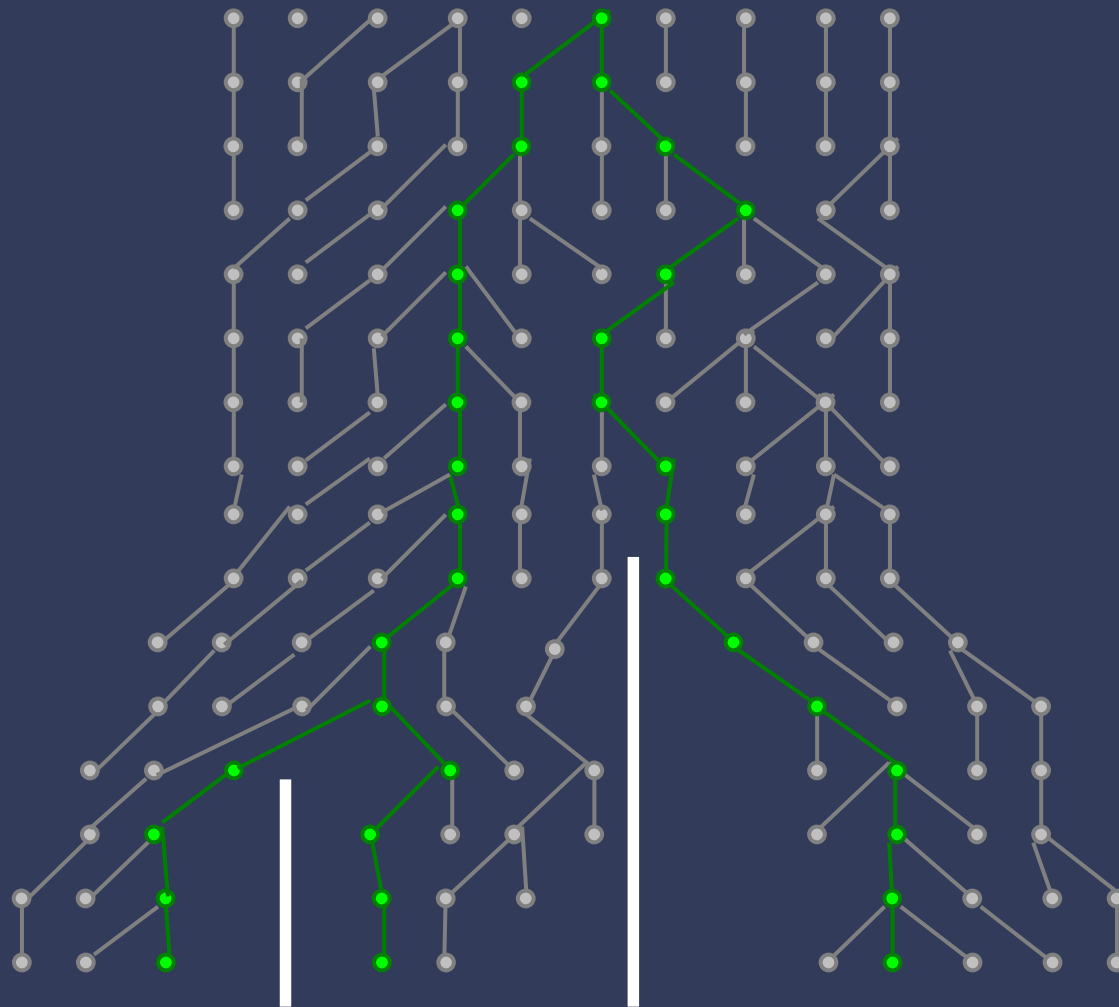
Populations



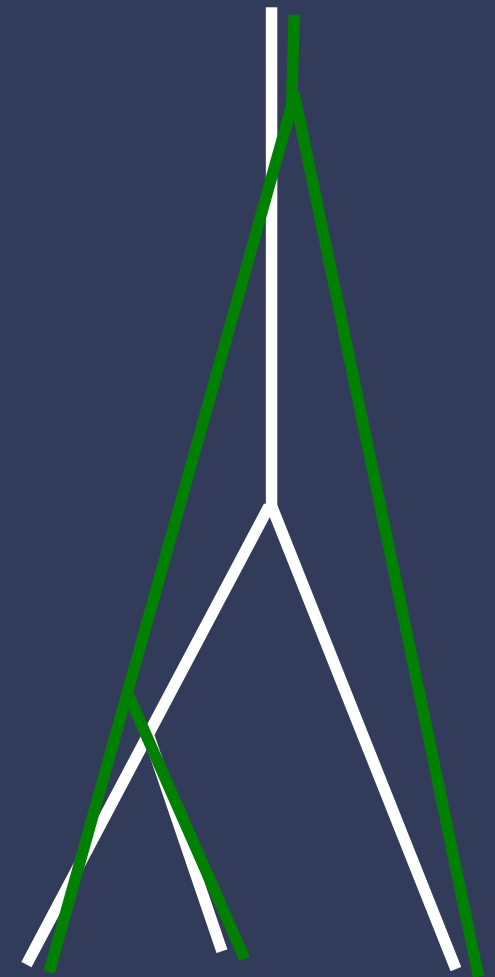
Populations and species



Individuals



Populations

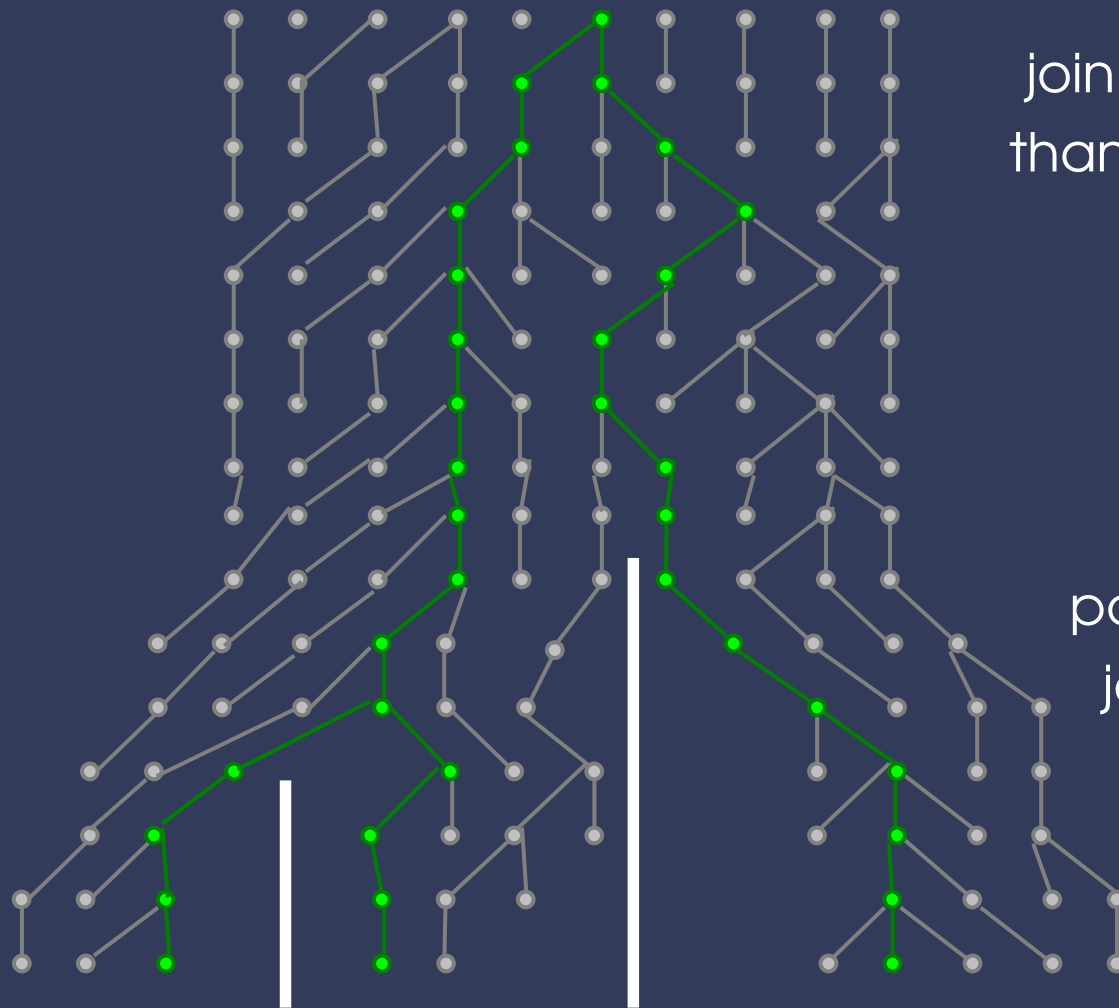


Populations and species



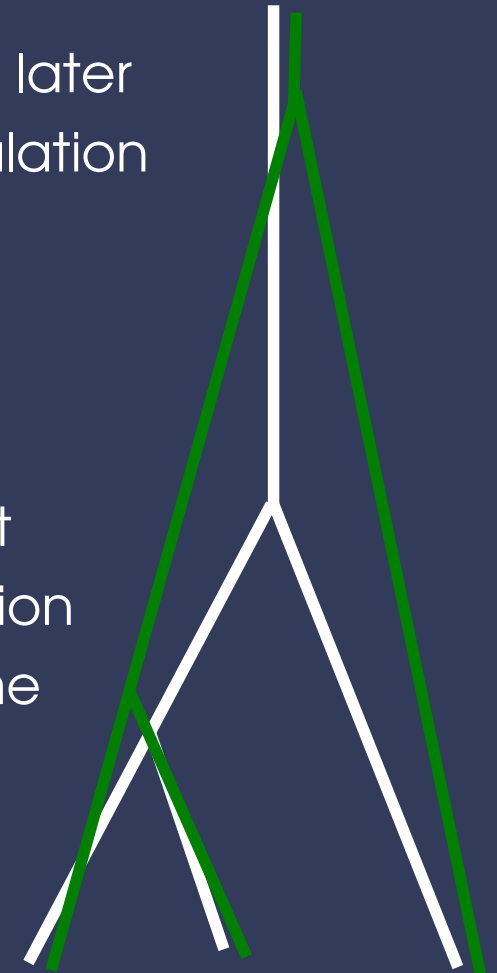
Individuals

Populations



join much later
than population

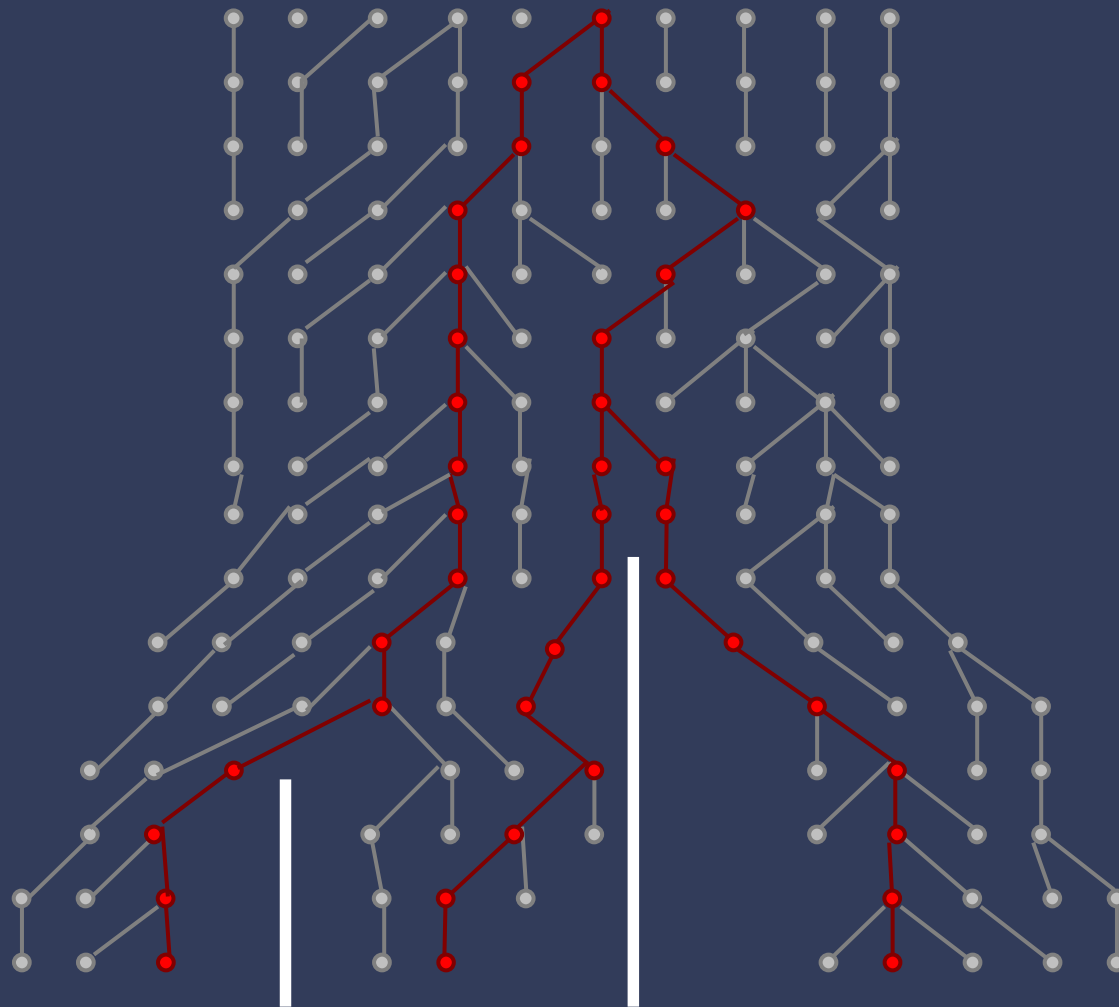
join at
population
join time



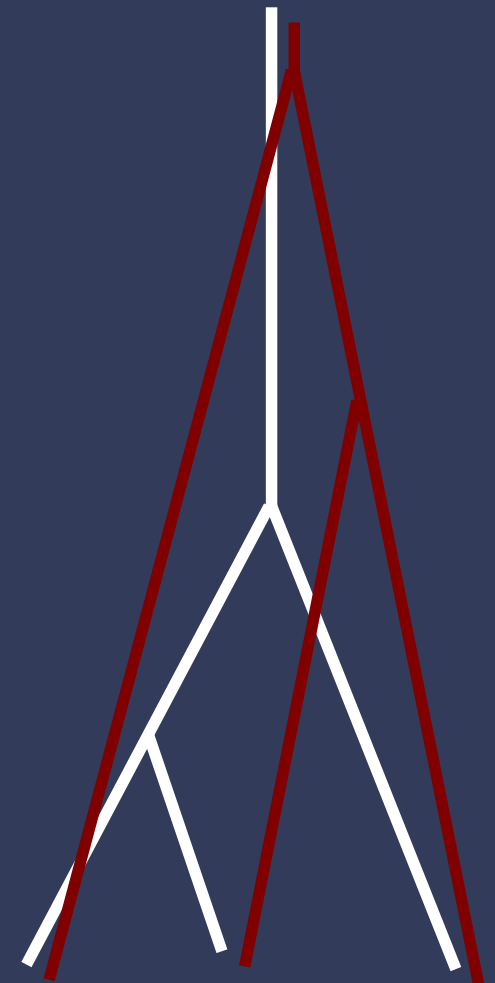
Populations and species



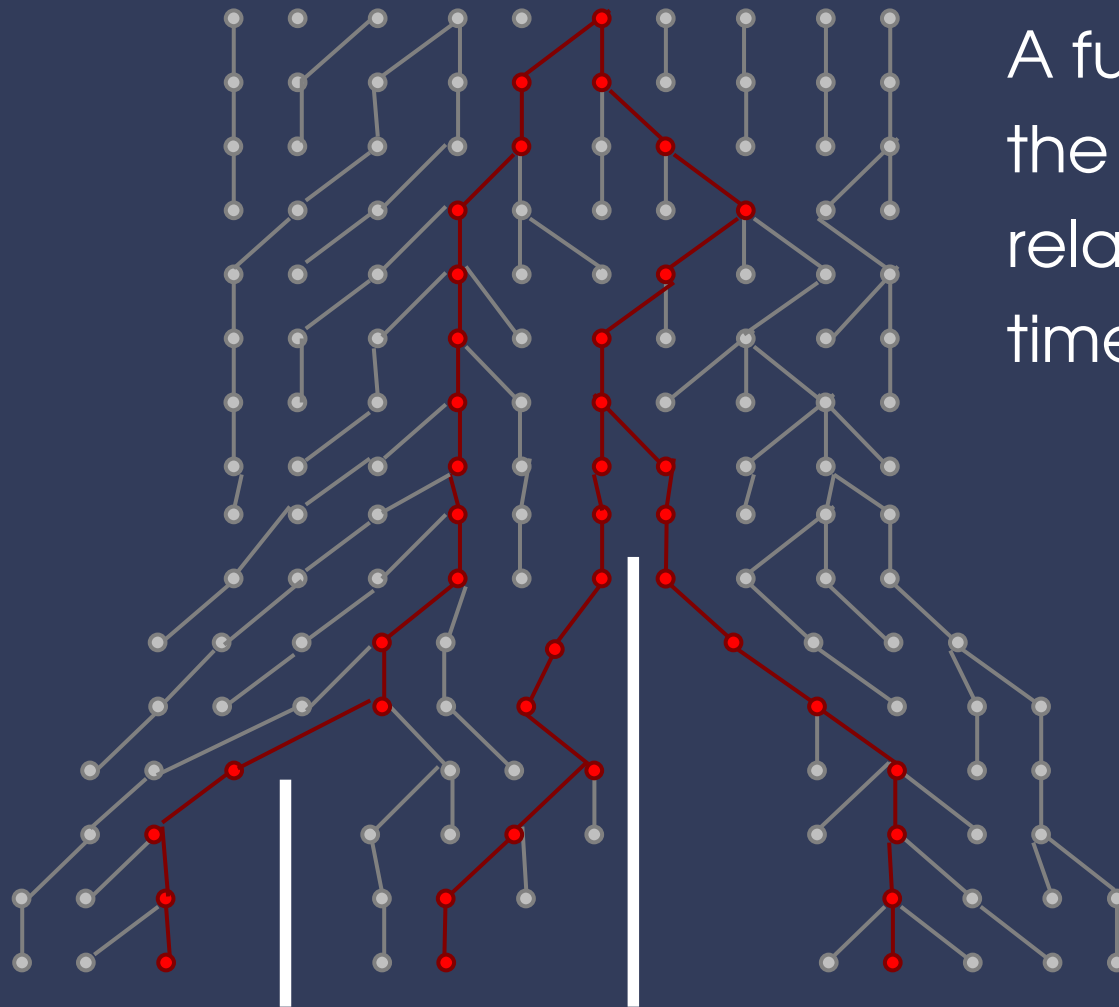
Individuals



Populations



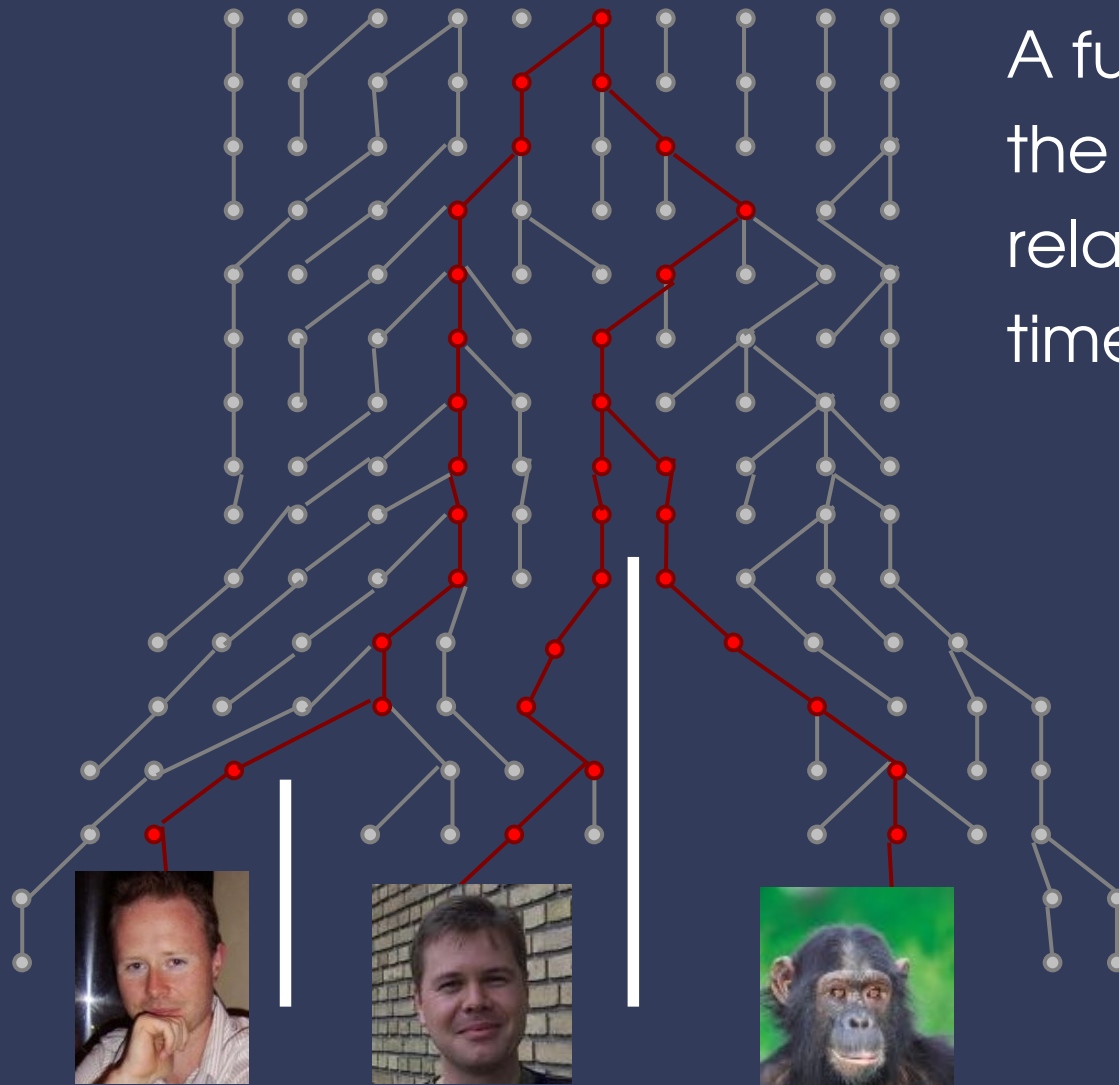
Populations and species



A funny thing happens if
the population size is large
relative to the splitting
time...

...for speciation, the
time is too long...

Populations and species



A funny thing happens if the population size is large relative to the splitting time...

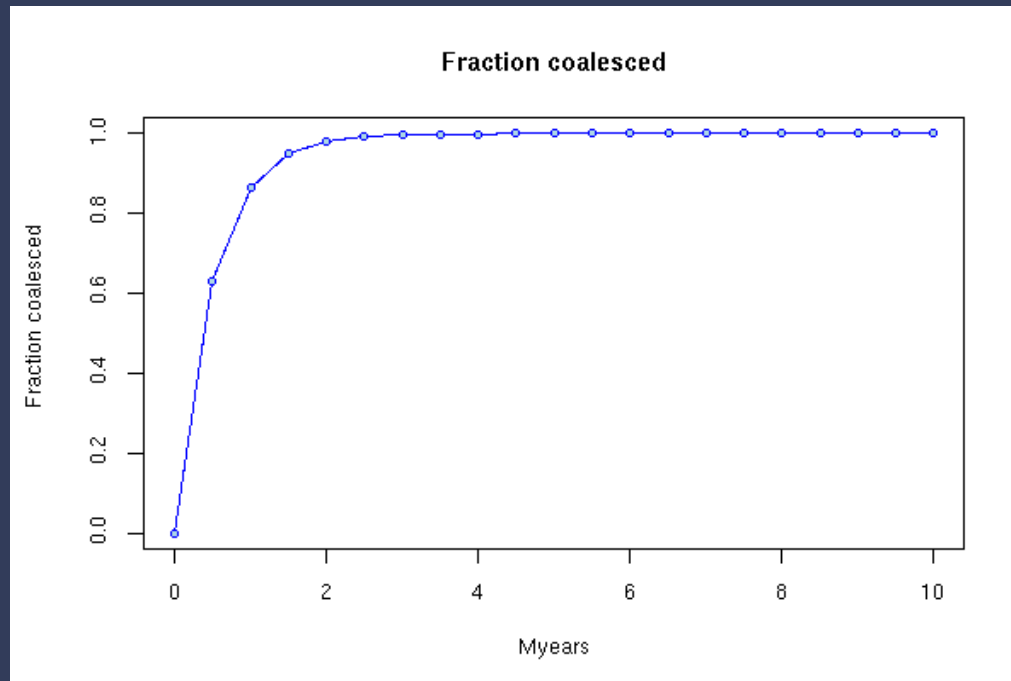
...for speciation, the time is too long...

...and no human is closer related to chimps than any other human!

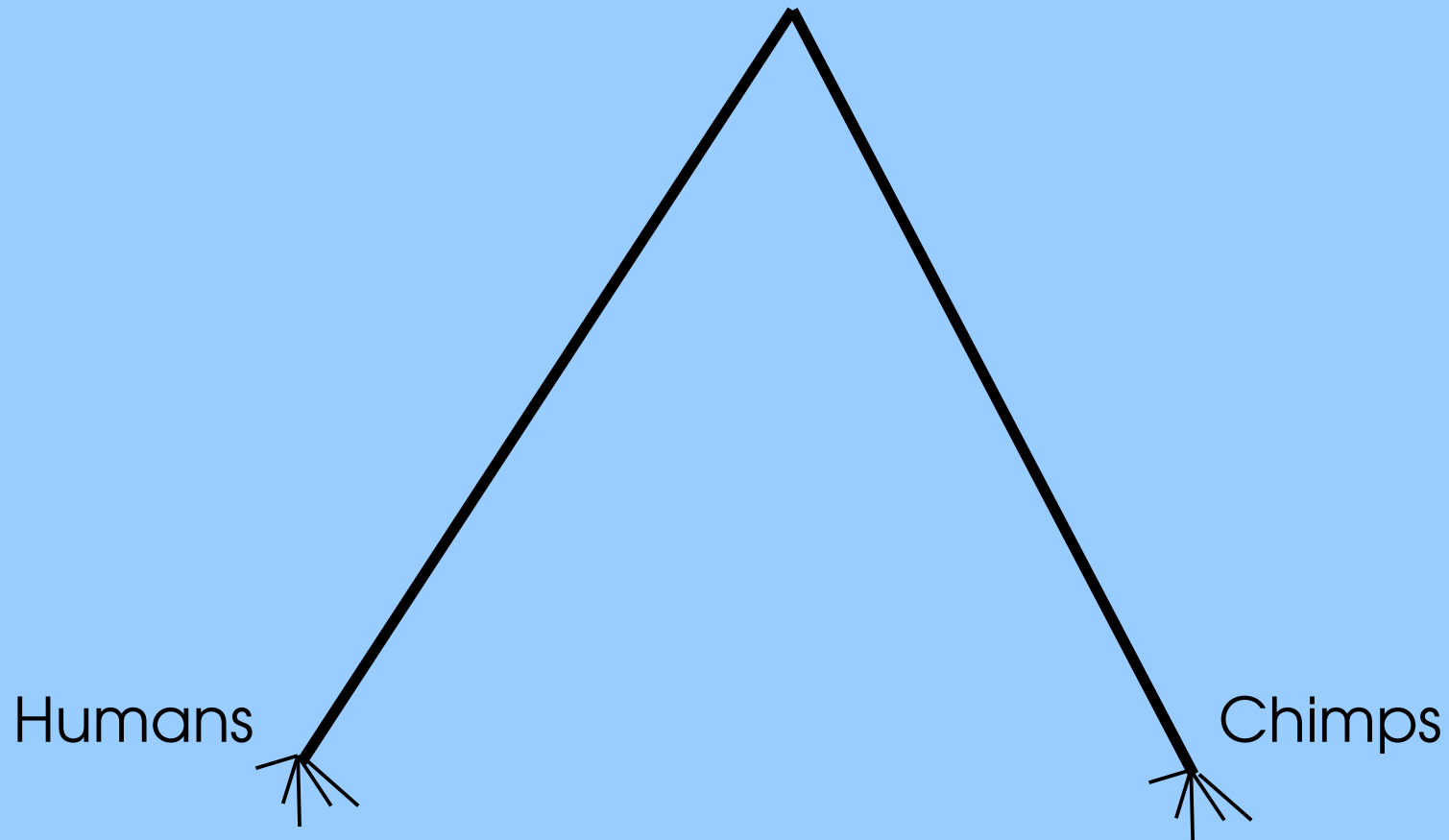
Populations and species



- The time for k lines to coalesce is distributed as $E(k(k-1)/2)$ in units of $2N_e$ generations.
 - Generation time ~ 25 years
 - N_e for humans ~ 10000



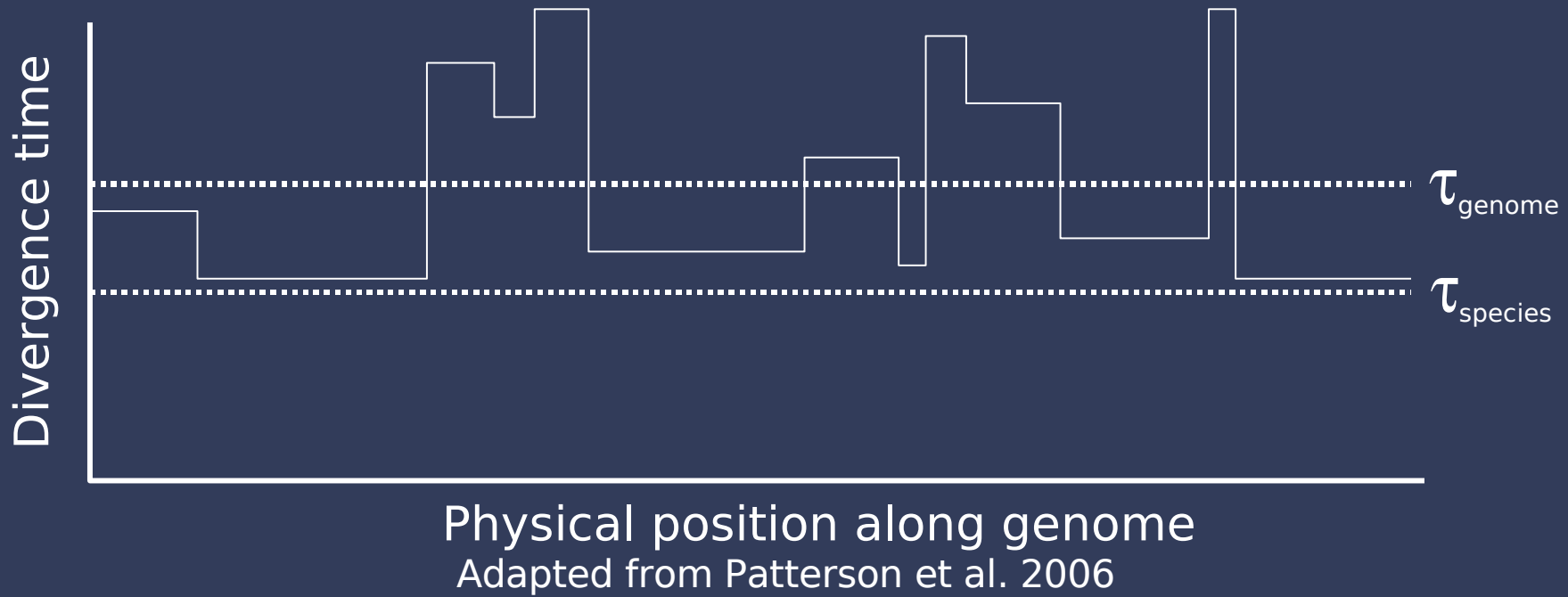
Populations and species



→ We will *all* have coalesced before we meet the chimps

- But there is ***recombination!***
 - Breaks up DNA in segments with separate histories
 - Single extant sequence reaches an equilibrium back in time
 - We ***will*** have several segments meeting the chimp!

Recombination



Species and segment trees



Case A:

Only Human and Chimp can coalesce here. Always consistent with species tree.



Case B:

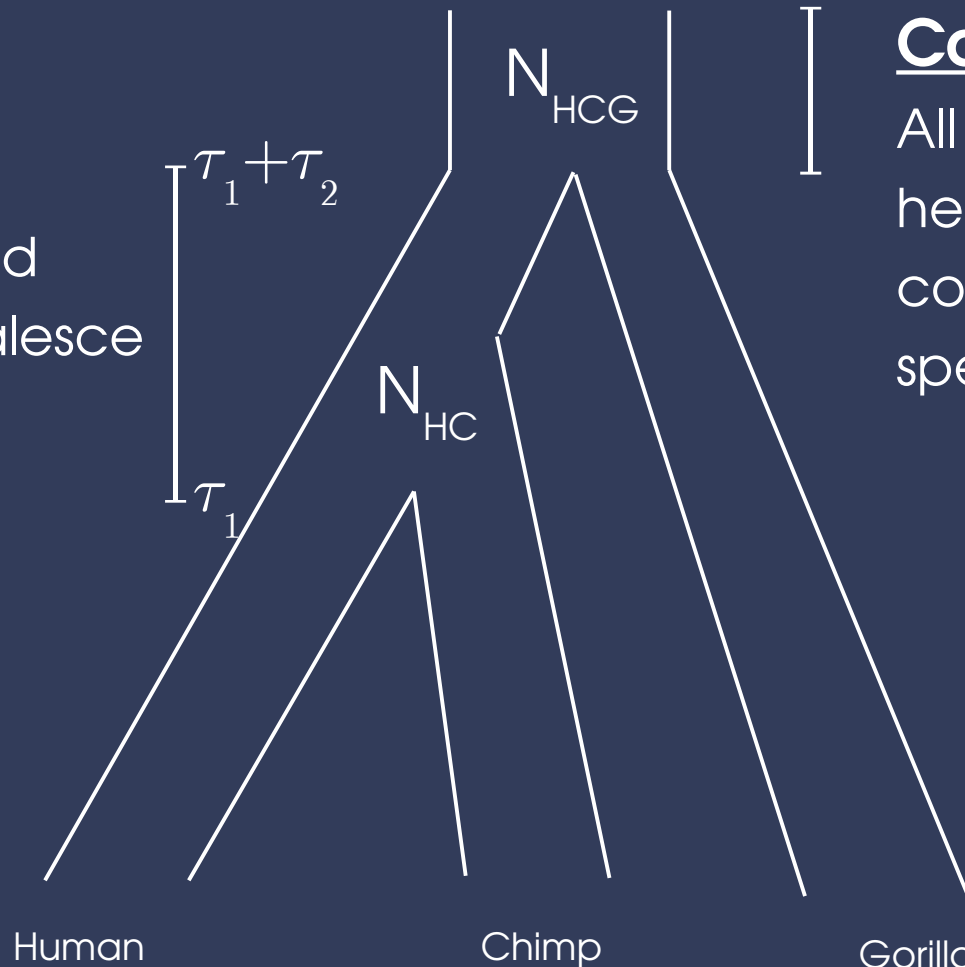
All species can coalesce here. Only a third consistent with species tree.

Species and segment trees



Case A:

Only Human and Chimp can coalesce here. Always consistent with species tree.



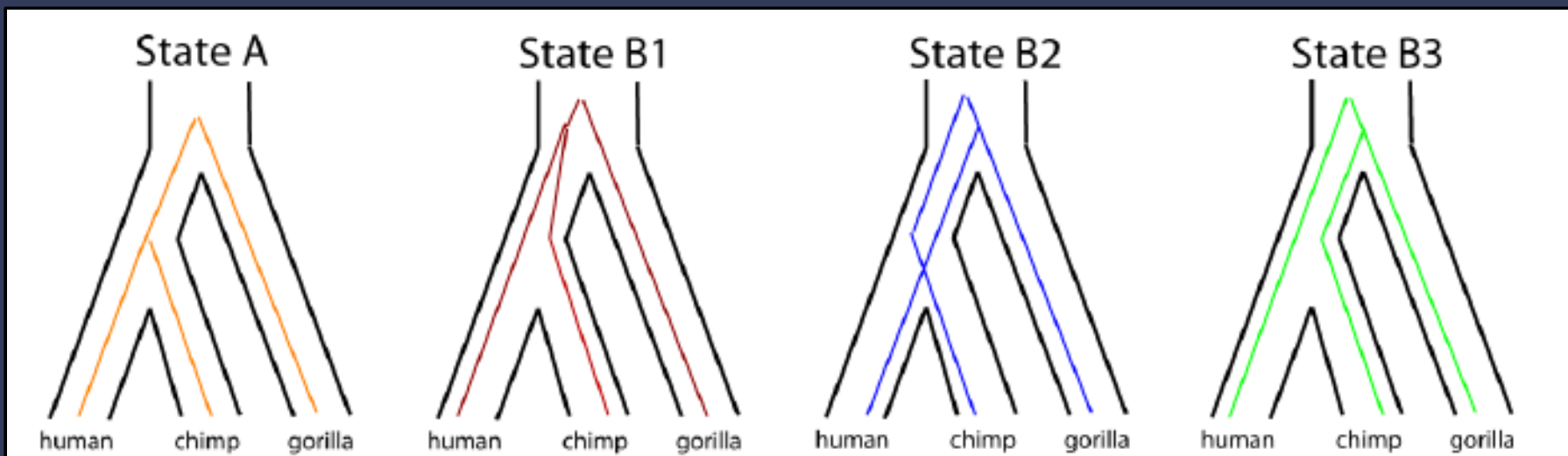
Case B:

All species can coalesce here. Only a third consistent with species tree.

We can get the probability of A or B based on split times and population size

Species and segment trees

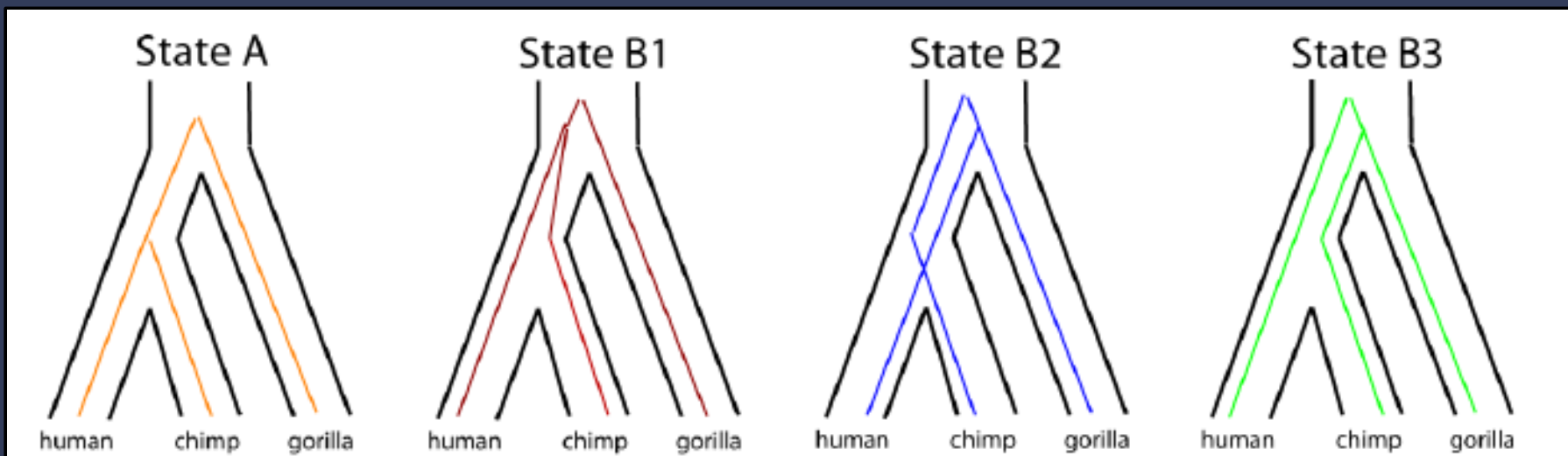
We can express the probability of either of these trees in terms of splitting times and effective population sizes...



Species and segment trees

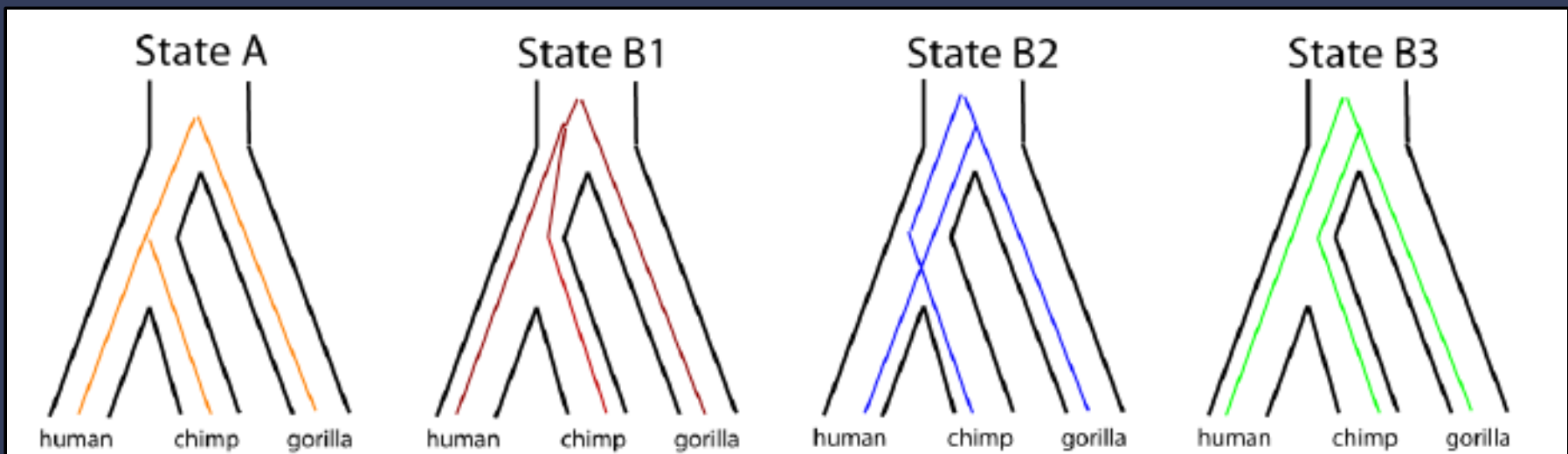
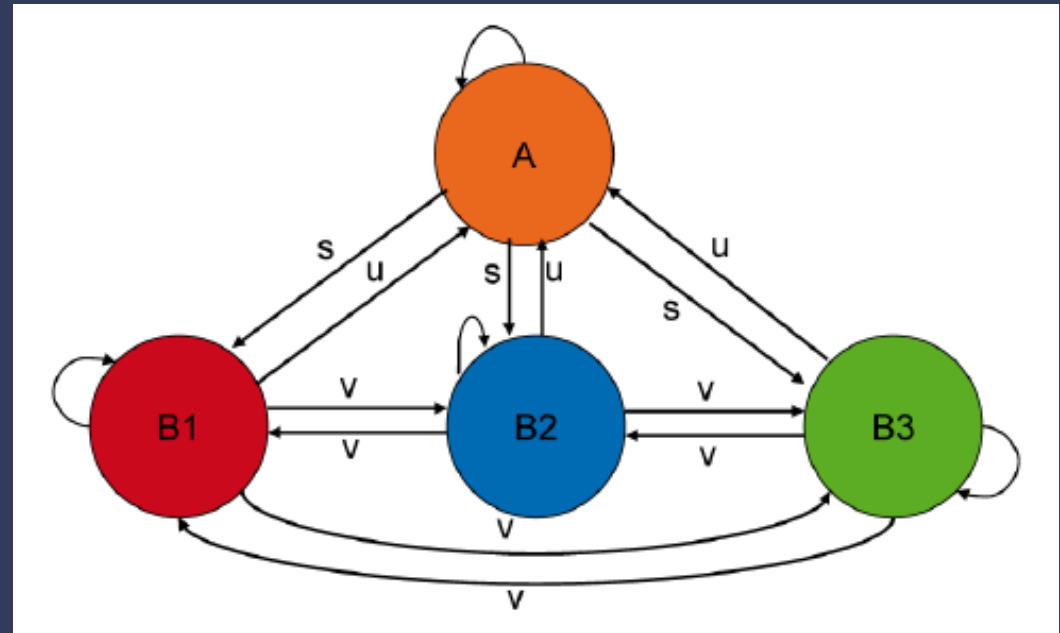
We can express the probability of either of these trees in terms of splitting times and effective population sizes...

...or alternatively get the time parameters from the tree probabilities.



Species and segment trees

We obtain the tree probabilities from an approximation to the coalescence process: a *hidden Markov model*

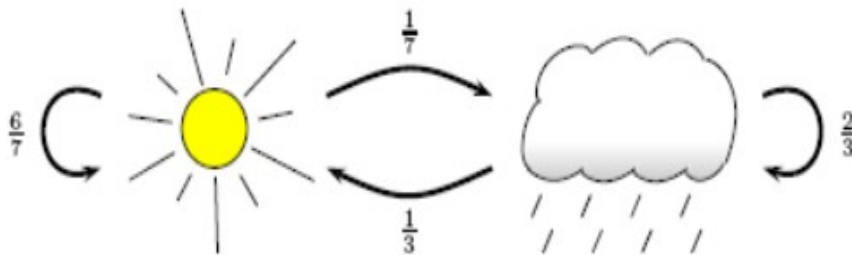


Markov models

A **Markov model** is an automaton where transitions are probabilistic, i.e. each transition to the next state is taken with a certain probability.

A **run** is a sequence of states generated by the model.

Model *M*:



A **run** emits a sequence of states:

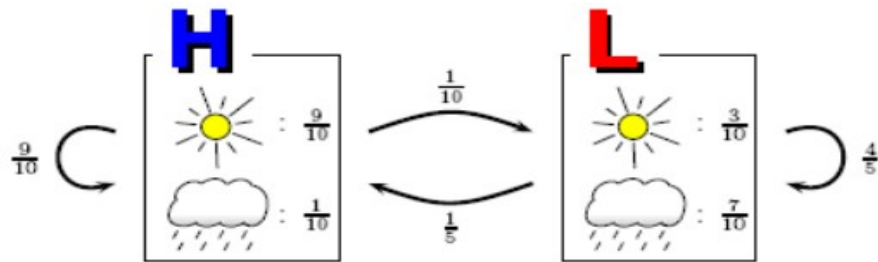


Hidden Markov models (HMMs)

A *hidden Markov model* in addition has emission symbols and probabilities. In each state it emits symbols with certain probabilities.

A *run* is a sequence of both states and emissions. We only *observe* the emissions.

Model *M*:



A *run* follows a sequence of states:

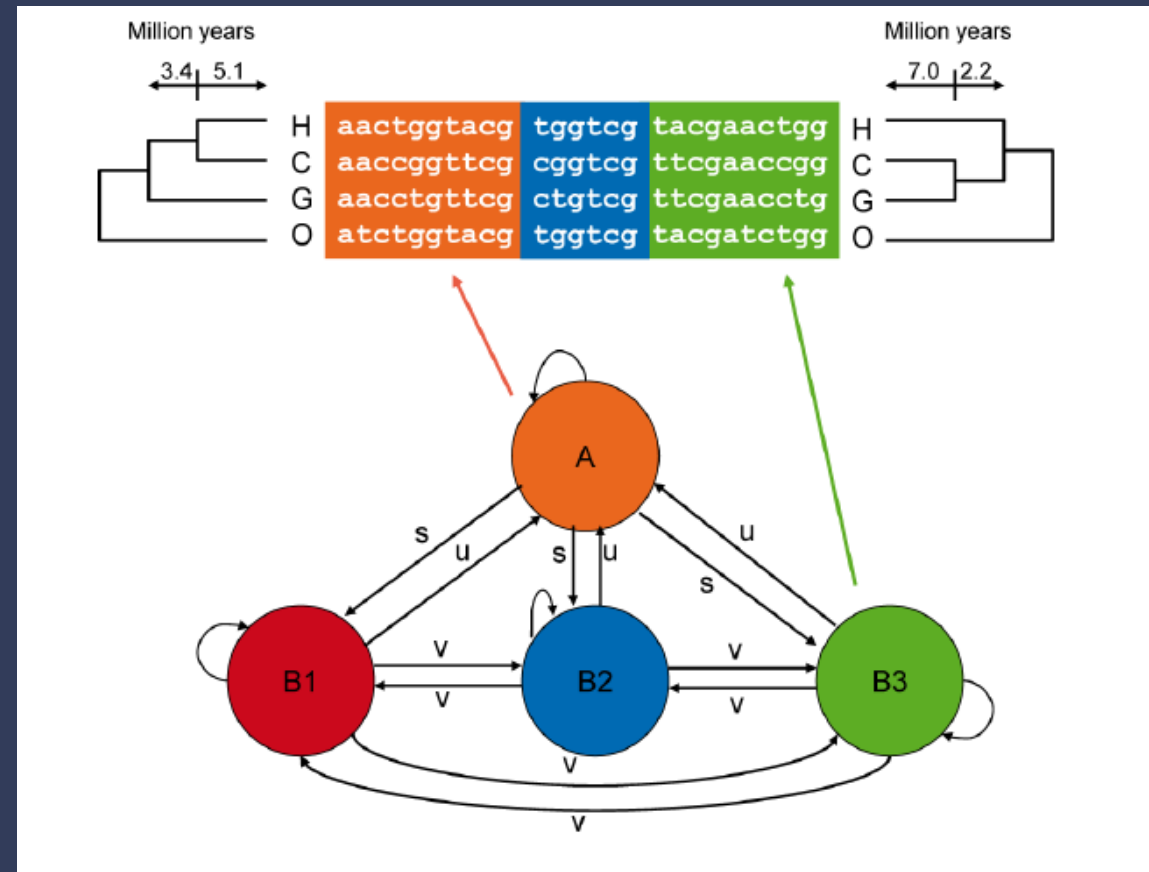
H H L L H

And *emits* a sequence of symbols:

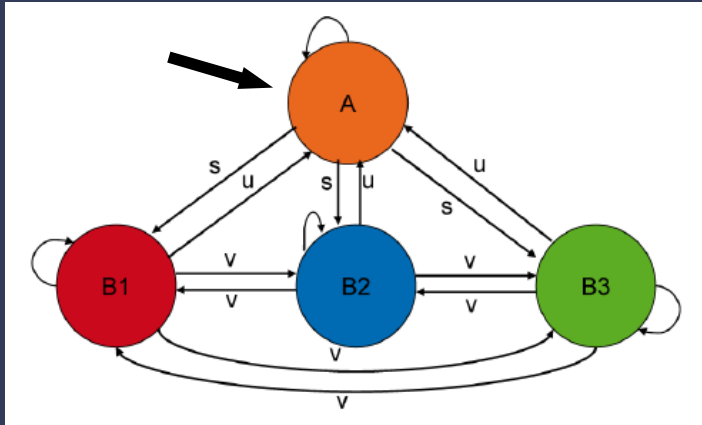


A coalescence HMM

- States correspond to the four trees
- Emission probabilities using dynamic programming algorithm
 - Branch lengths are mean branch lengths for each tree type

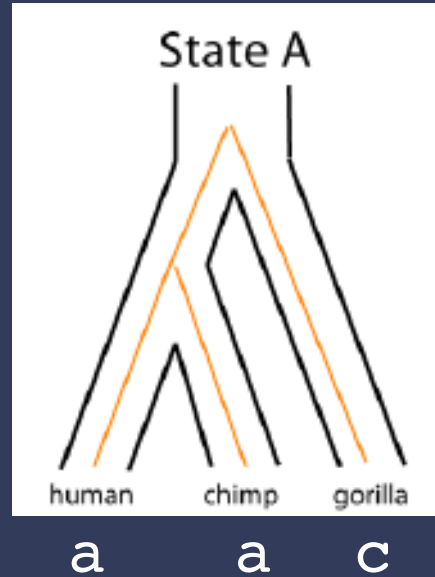
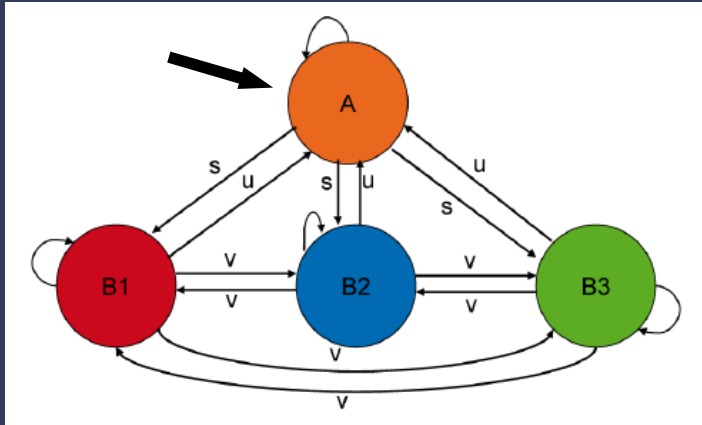


A coalescence HMM



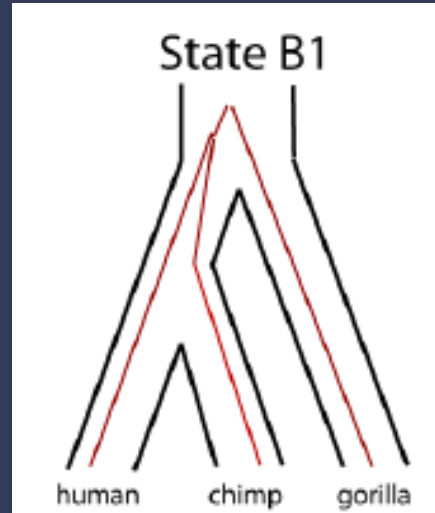
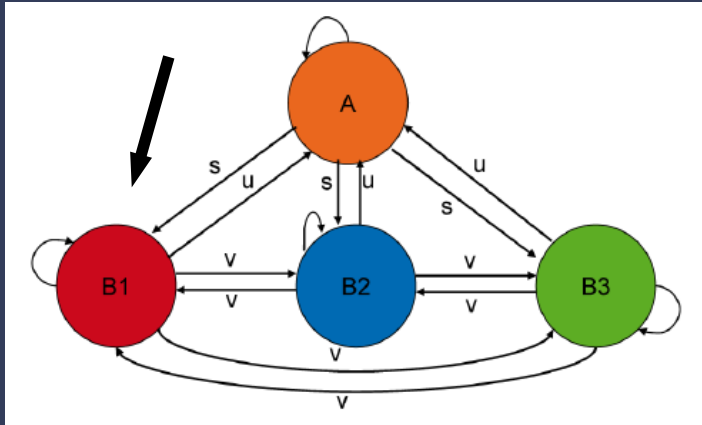
A

A coalescence HMM



A
a
a
c

A coalescence HMM



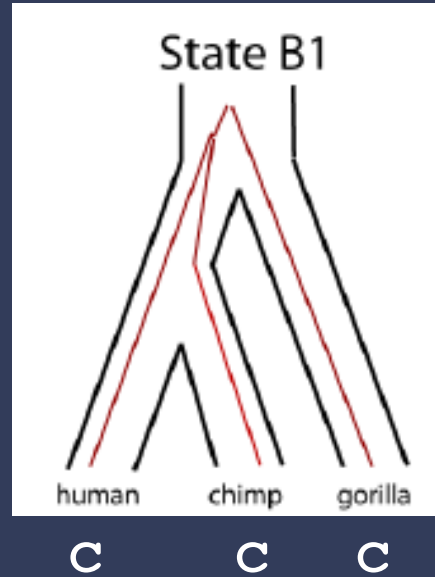
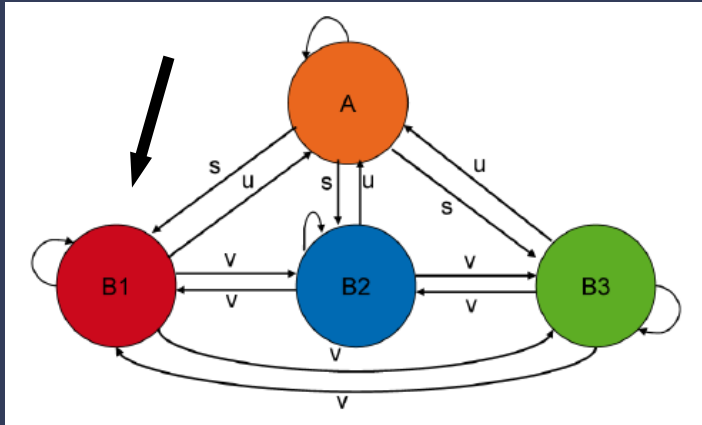
A B1

a

a

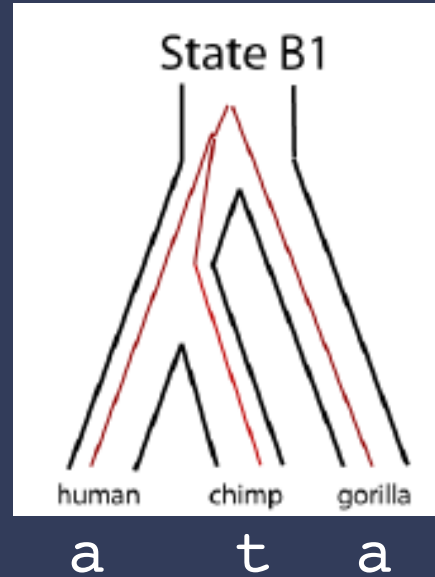
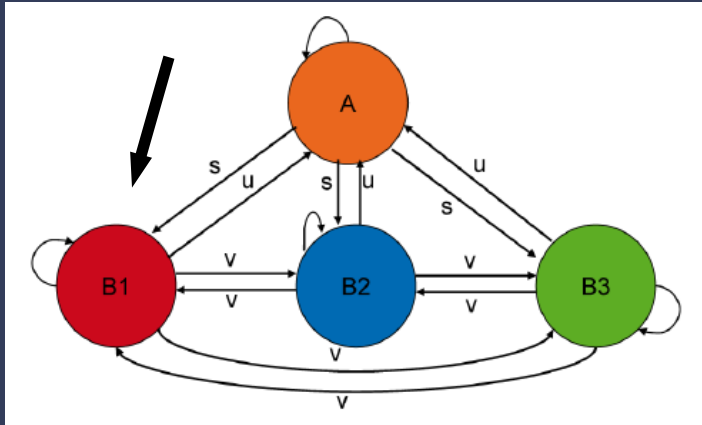
c

A coalescence HMM



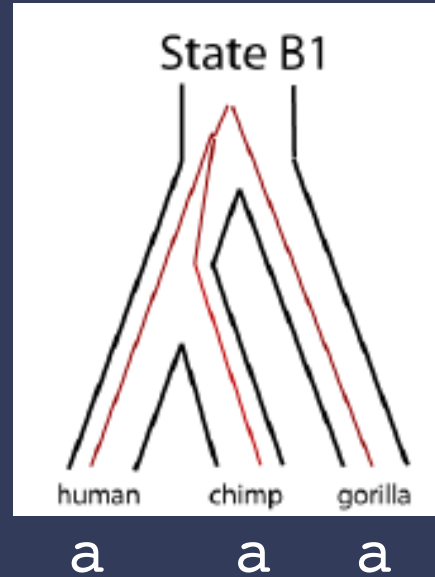
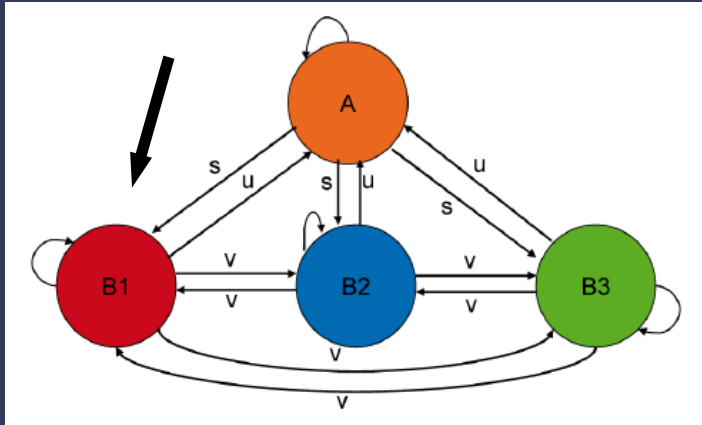
A	B1
a	c
a	c
c	c

A coalescence HMM



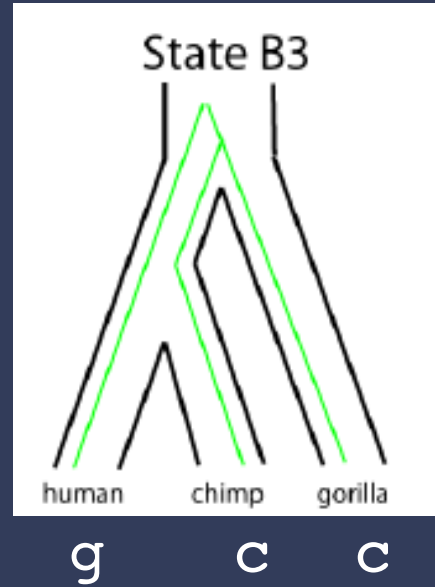
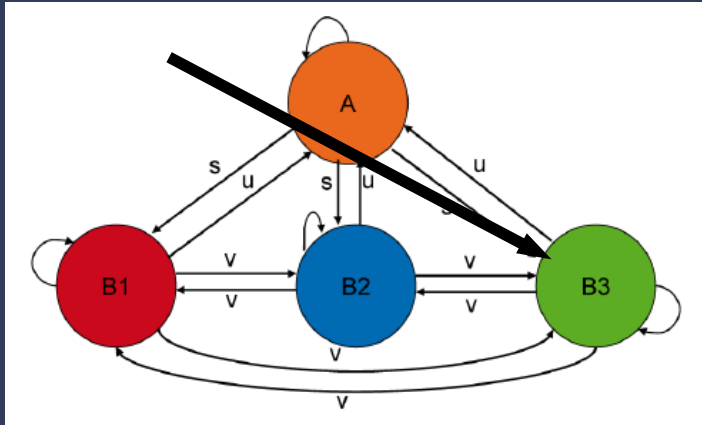
A	B1	B1
a	c	a
a	c	t
c	c	a

A coalescence HMM



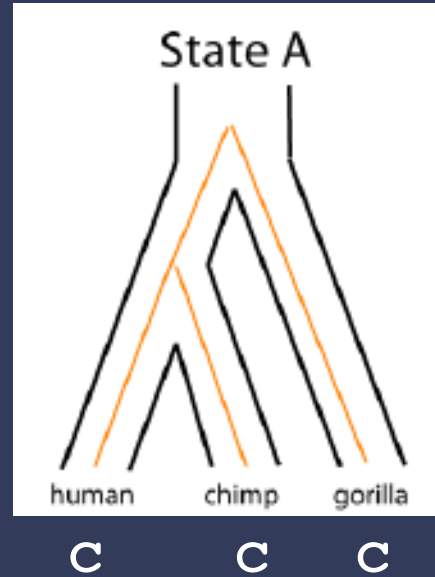
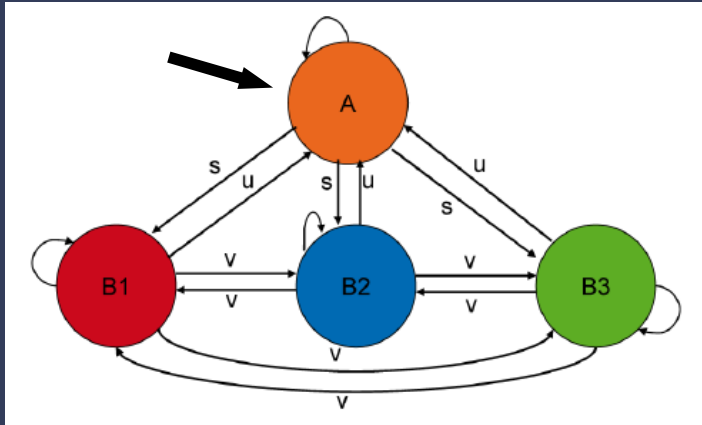
A	B1	B1	B1
a	c	a	a
a	c	t	a
c	c	a	a

A coalescence HMM



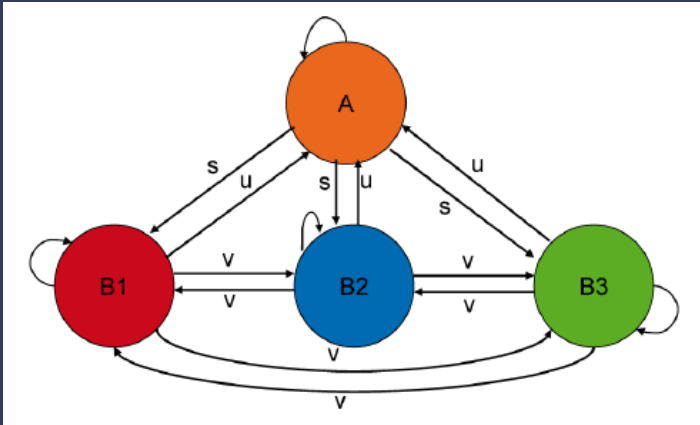
A	B1	B1	B1	B3
a	c	a	a	g
a	c	t	a	c
c	c	a	a	c

A coalescence HMM



A	B1	B1	B1	B3	B3	B3	A	A	A
a	c	a	a	g	a	t	c	a	c
a	c	t	a	c	t	c	t	a	c
c	c	a	a	c	t	c	t	a	c

A coalescence HMM



$$p(\mathbf{s}, \mathbf{e} \mid \tau, \epsilon) = \prod_i \epsilon(e_i \mid s_i) \tau(s_i, s_{i+1})$$

$$p(\mathbf{e} \mid \tau, \epsilon) = \sum_{\mathbf{s}} \prod_i \epsilon(e_i \mid s_i) \tau(s_i, s_{i+1})$$

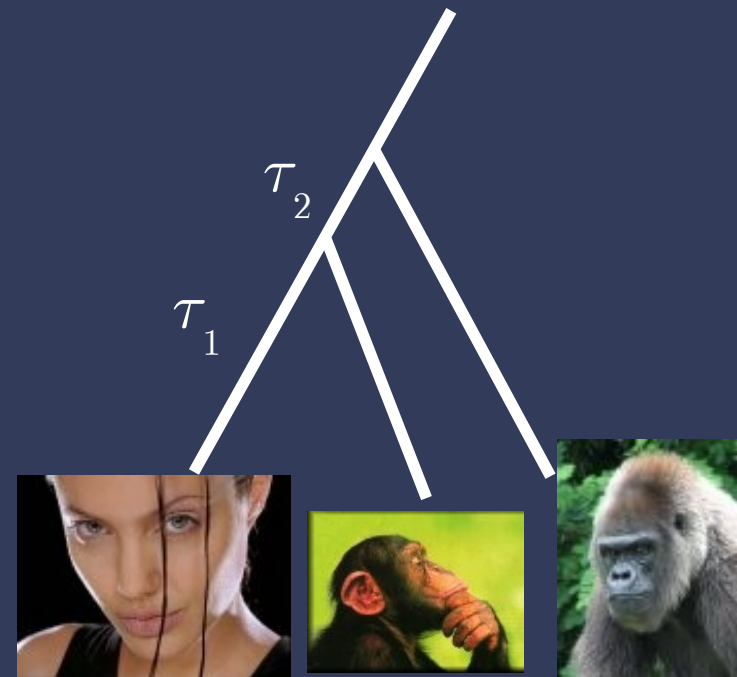
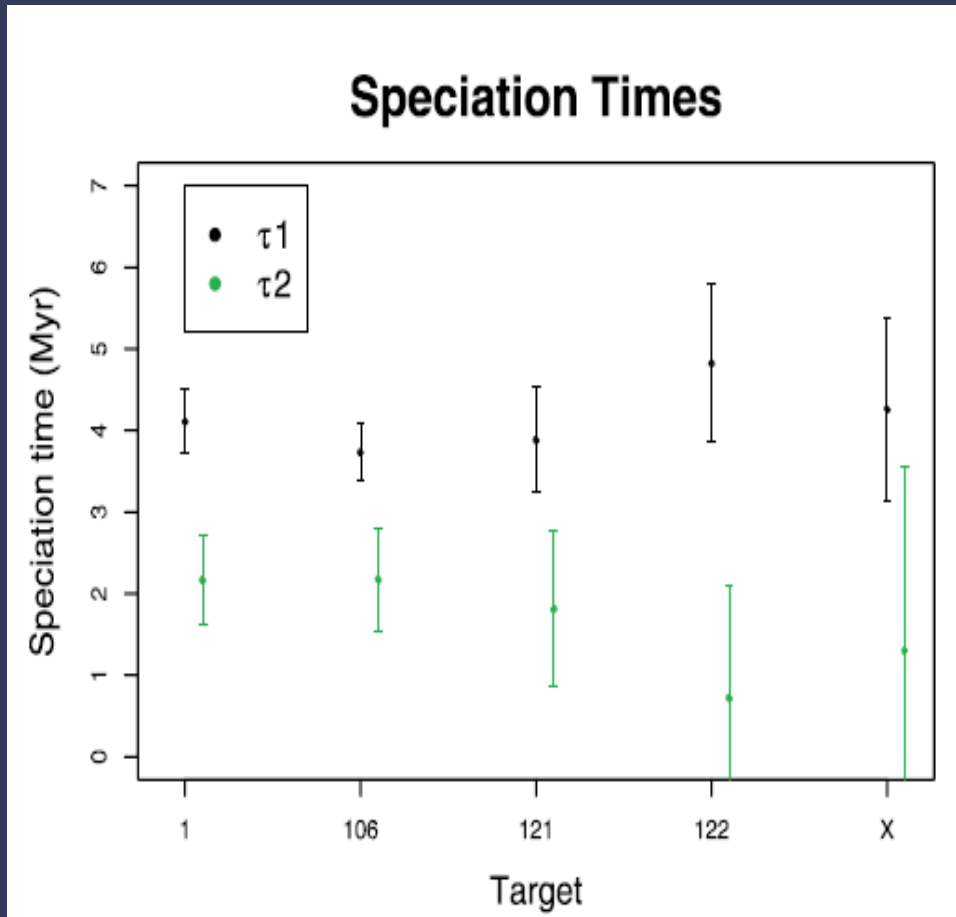
A	B1	B1	B1	B3	B3	B3	A	A	A
a	c	a	a	g	a	t	c	a	c
a	c	t	a	c	t	c	t	a	c
c	c	a	a	c	t	c	t	a	c

Inference in the CoalHMM



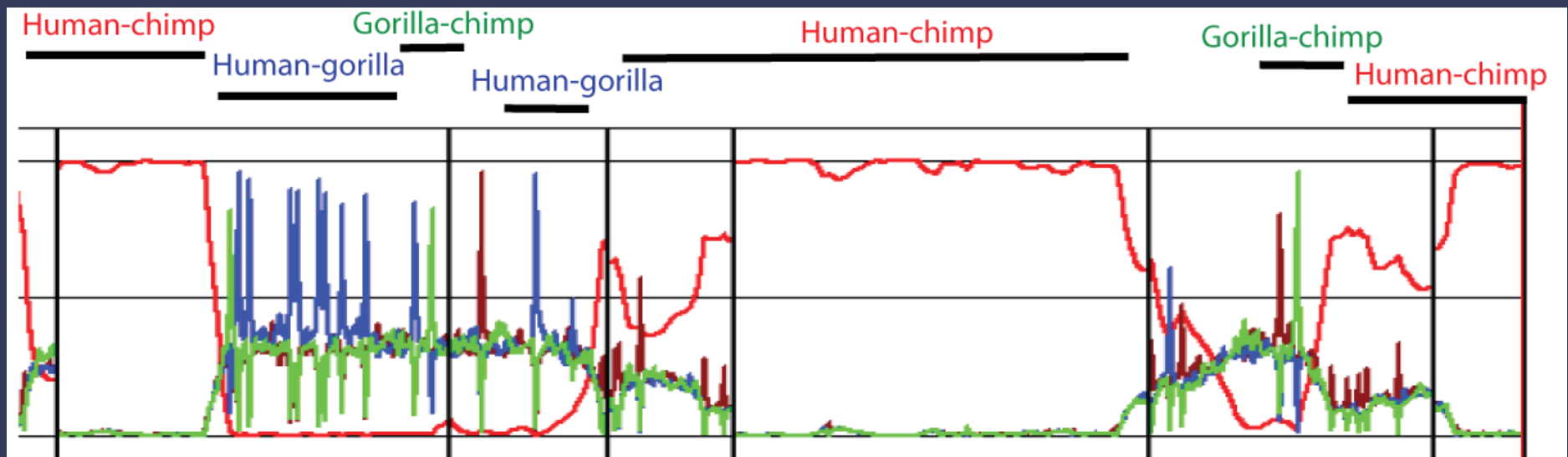
- We use standard algorithms for estimating the HMM parameters
 - Transition probabilities gives us:
 - State probabilities, which gives us:
 - Coalescence process parameters, which gives us:
 - ***Speciation times!***

Results



Annotation in the CoalHMM

Posterior probabilities of each tree (probability of being in a particular tree given the sequence data) along a genomic region:



Summary



- Dating the speciation between humans, chimps, and gorilla
 - Expressed as a molecular evolution / population genetics problem
 - Approximated by a machine learning / statistical model (HMM)
 - Estimated HMM parameters gives us speciation times
- Puts speciation time of human-chimp ~4 My ago and (human-chimp) – gorilla ~5-6 My ago

The end



Thank you!

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1802818&rendertype=abstract>

Cheers!



See you in the bar!