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

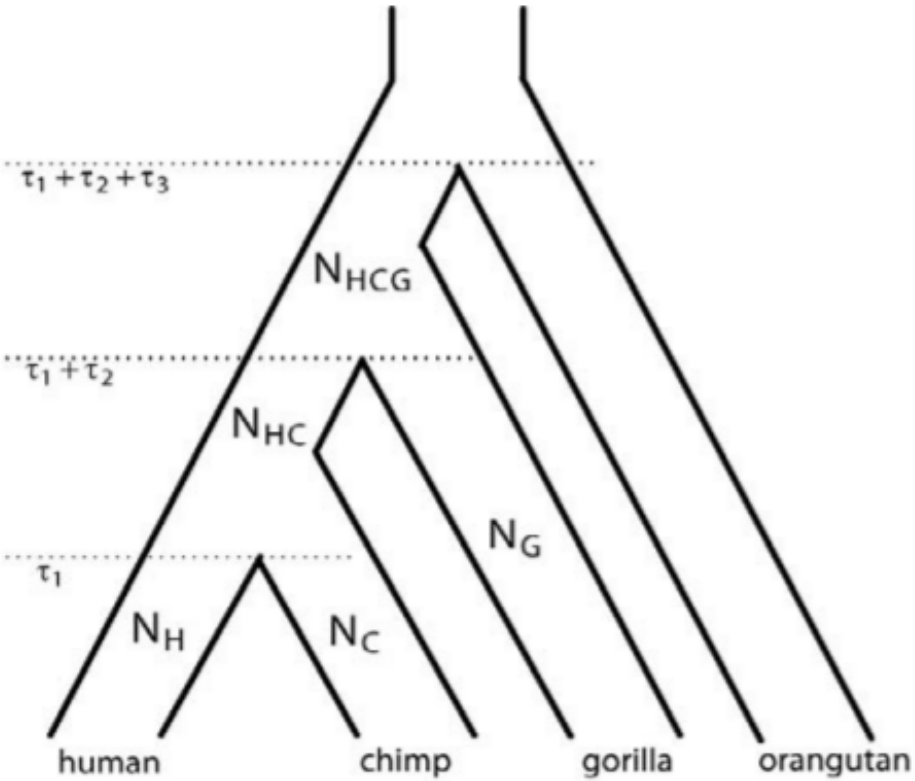
Asger Hobolth, Ole F. Christensen, Thomas Mailund, Mikkel H. Schierup.

Fiona Grimson

30 April 2013

Scientific Motivation

Learn about the evolutionary history of Human, Chimpanzee, Gorilla and Orangutan by comparing their DNA sequences.



Does " Millenium Man", dated to 6 million years ago, belong to the human lineage or the human-chimp lineage?

Scientific Motivation

The Holy Grail: Ancestral Recombination Graph (ARG)

- Record of every recombination and coalecent event, at every locus.
- Large, complicated and not Markov when viewed as a process along the genome.
- There are methods to estimate it, but they are only feasible for a short piece of genome.

Scientific Motivation

The Holy Grail: Ancestral Recombination Graph (ARG)

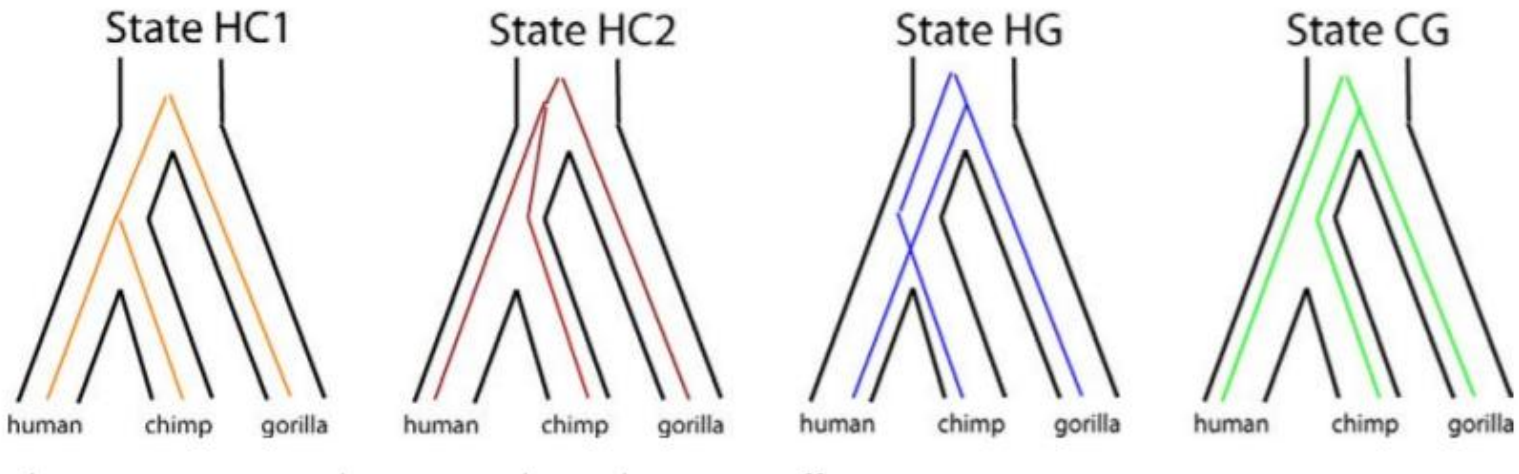
- Record of every recombination and coalecent event, at every locus.
- Large, complicated and not Markov when viewed as a process along the genome.
- There are methods to estimate it, but they are only feasible for a short piece of genome.

The Approximation: Coalescent Hidden Markov Model

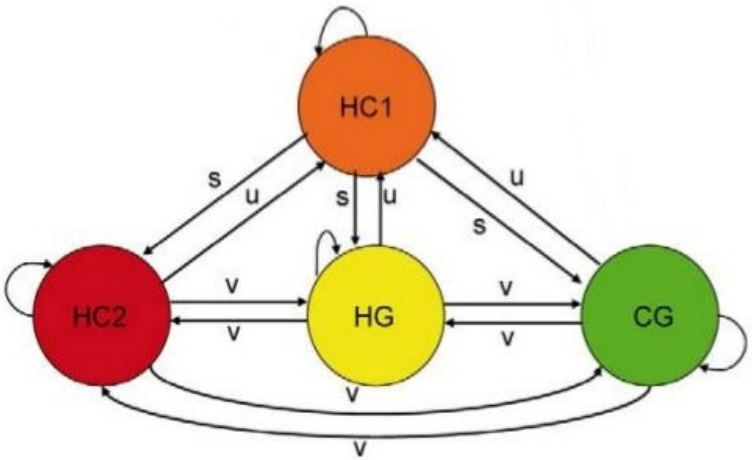
- Small, simple, and Markov model that can be applied to long chunks of genome.

The Hidden States

There are (infinitely) many possible genealogies consitent with the species tree. In this paper, coalescents are summarised by four topologies,



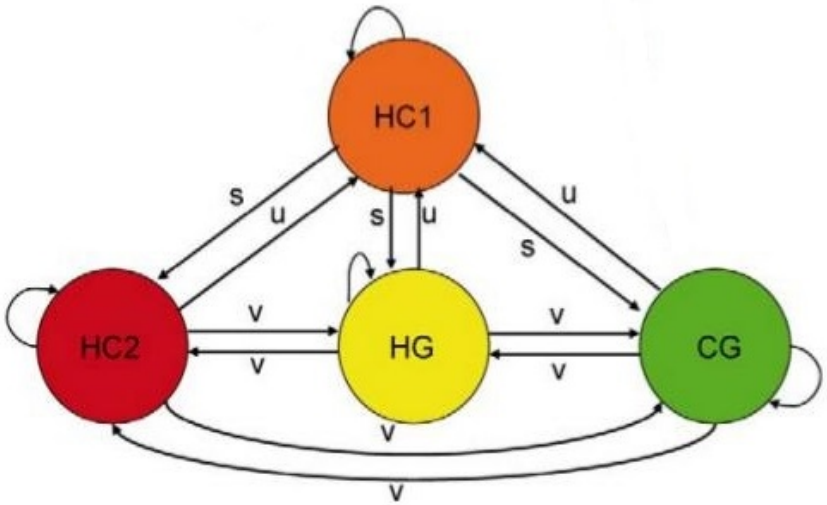
and move between the states when there is a recombination event



The Hidden States

This process is not observed directly! But we assume it is a discrete time Markov chain along the sequence.
The transition probaility matrix is:

$$\mathbf{P} = \begin{bmatrix} . & s & s & s \\ u & . & v & v \\ u & v & . & v \\ u & v & v & . \end{bmatrix}$$

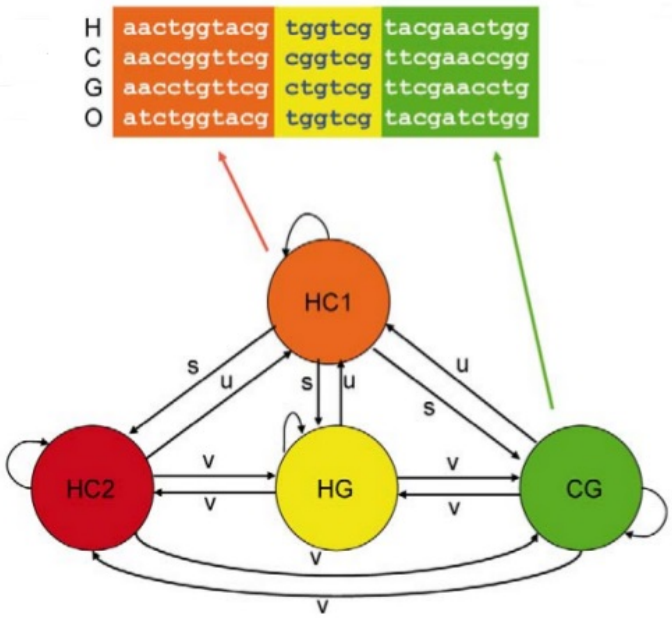


Notation: ϕ_i is the hidden state at a locus i.

- Transition Probability $P(\phi_{i-1}, \phi_i) = \mathbf{P}[\phi_{i-1}, \phi_i]$
- Stationary Distribution $\Psi = \left(\frac{u}{3s+u}, \frac{s}{3s+u}, \frac{s}{3s+u}, \frac{s}{3s+u}\right)$

The Emissions

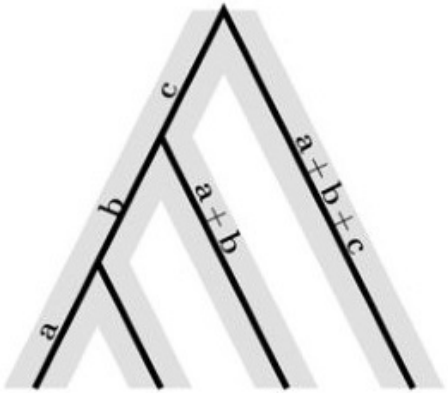
We observe the genome sequences, which are emissions from these hidden states:



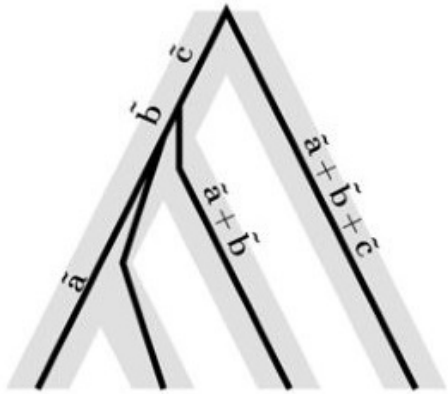
The emission process is a continuous time Markov Chain that models the evolution of a particular DNA base over evolutionary time.

We want to calculate emission probabilities.

The Emissions



Branch lengths for the standard genealogy $HC1$



Branch lengths for the alternative genealogies $HC2, HG, CG$.

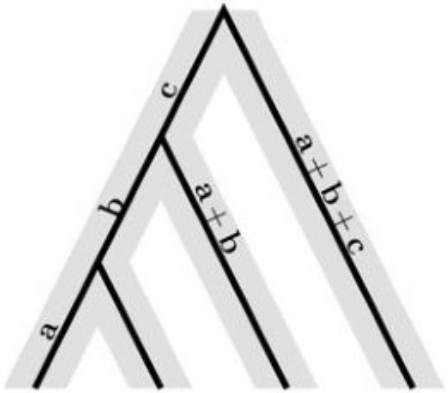
- Times are the branch lengths.
- a, b, c for HC1 topology, $\tilde{a}, \tilde{b}, \tilde{c}$ for HC2, HG, CG topologies.
- Coalescent theory tells us that

$$a + b + c = \tilde{a} + \tilde{b} + \tilde{c}$$

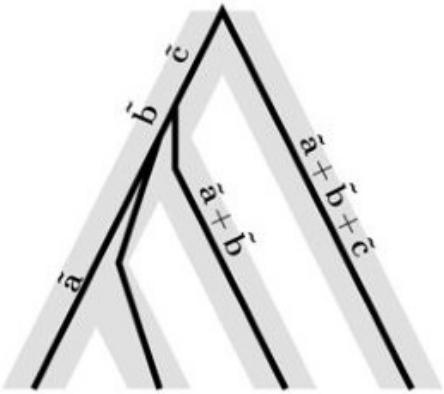
$$\tilde{b} = 1.5(a + b - \tilde{a})$$

so $a, b, c, \tilde{a}$ are the free parameters.

The Emissions



Branch lengths for the standard genealogy $HC1$

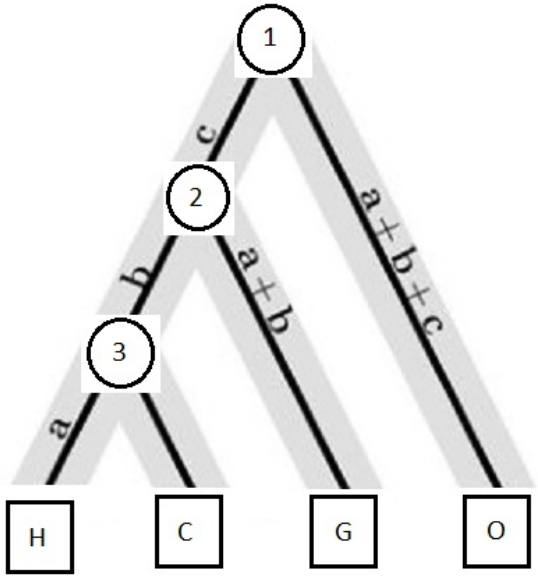


Branch lengths for the alternative genealogies $HC2, HG, CG$.

- Rates are given by 4x4 matrix $\mathbf{Q}$, rate of A, T, C, G substitutions.
- Many suggestions in literature of resonable $\mathbf{Q}$ matrices
- In the paper $\mathbf{Q}$ has a parametric form known as strand symmetric. The parameters are estimated once, and then treated as known.
- Choice of $\mathbf{Q}$ doesn't matter much
- Process is reversible and starts at its stationary distribution $\boldsymbol{\pi}$.
  P(transition) = exp(Q $\times$ time)

The Emissions

So how to calculate emission probability?



- Hidden state tells us what graph to use
- Use law of total probability to sum transition probabilities between nodes.

$$\text{P(Data}|\text{ Hidden State)} = \sum_{Node1} \sum_{Node2} \sum_{Node3} P(1)P(1 \to Orang)$$
$$P(1 \to 2)P(2 \to Gor.)P(2 \to 3)$$
$$P(3 \to Chimp.)P(3 \to Human).$$

The Parameters

The free parameters are $\eta = (s, u, v, a, b, c, \tilde{a})$

We estimate the parameters by maximum likelihood.

- $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_L]$ alignment at L sites
- $\phi = (\phi_1, \phi_2, ..., \phi_L)$ where $\phi_i \in \{HC1, HC2, HG, CG\}$
- Ψ_η Initial distribution (stationary) of hidden states

Joint probability of alignment and path through hidden states:

$$P(\mathbf{X}, \phi|\eta) = \Psi_\eta(\phi_1) P_\eta(\mathbf{X}_1|\phi_1) \prod_{i=2}^{L} P_\eta(\phi_{i-1}, \phi_i) P_\eta(\mathbf{X}_i|\phi_i)$$

Likelihood:

$$L(\eta) = \sum_{\mathbf{\Phi}} P(\mathbf{X}, \phi|\eta)$$

Estimating Parameters

Maximise the likelihood (and avoid summing over 4^L paths) by using an EM algorithm.

1. Set Q
2. Initialise η
3. Find most likely path ϕ given η
   use Viterbi algorithm
4. Find $\eta = \text{argmax } P(\mathbf{X}, \phi|\eta)$ given ϕ
   use Newton-Rapshon or Baum-Welch algorithm
5. Repeat (3) and (4) till convergence

Estimating Parameters

Maximisation Step is slow...

Newton-Raphson

- Move around 7D likelihood space by gradient.
- Slow

Estimating Parameters

Maximisation Step is slow...

Newton-Raphson

- Move around 7D likelihood space by gradient.
- Slow

Baum-Welch

- Specialised for HMM
- Fast - especially existing packages
- Finds closed forms for maxima

Estimating Parameters

Maximisation Step is slow...

Newton-Raphson

- Move around 7D likelihood space by gradient.
- Slow

Baum-Welch

- Specialised for HMM
- Fast - especially existing packages
- Finds closed forms for maxima
- Most implementations don't allow parameterised **P** or emission probability functions.
- My own code?
  Closed forms for s, u, v? Just $a, b, c, \tilde{a}$ by Newton-Raphson.

Testing The Model

Simulate Data

1. From the model.
   Are our estimates biased?
2. From a process that is closer to "real life".
   Is the model a reasonable approximation to the ARG?
   How does it compare with other methods?

Real Data

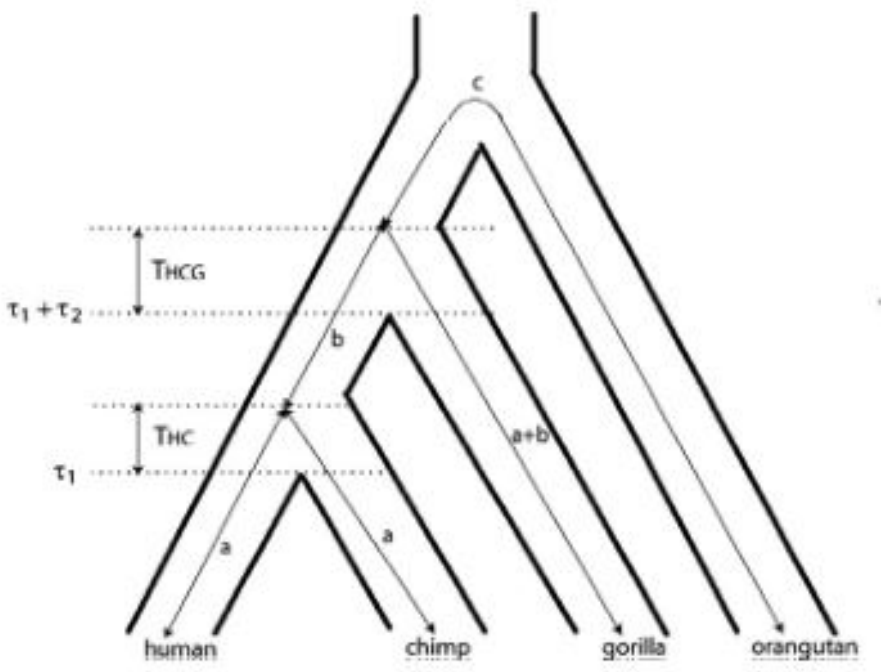
Target	Base Pairs
Target 1	1,255,492
Target 106	257,420
Target 121	230,666
Target 122	92,240

Conclusions

- Replication of Results?
  - Make maximisation step faster
  - Test the code
  - Get a better $\mathbf{Q}$
  - Simulate data
- What I haven't told you
  - How do the HMM parameters answer the scientific question?
  - What is the answer to the scientific question?
  - Criticisms of the model.

The Parameters

Want to know speciation times τ_1 and τ_2, and ancestral population sizes N_{HCG} and N_{HC}, which are functions of η



Key Ideas:

- Coalescent branch lengths exponentially distributed: $a, b, c, \tilde{a}, \tilde{b}, \tilde{c}$ are the means.
- Restrictions from species tree and hidden state topologies
- $T_{HCG} \sim \text{Exponential}(2N_{HCG}\mu)$
  $T_{HC} \sim \text{Exponential}(2N_{HC}\mu)$

The length of time in each state gives us the recombination rate.

Results

20 simulations:

Parameter	Mean	SD	True Value
τ_1	3.85	0.97	4
τ_2	1.58	1.47	1.5
N_{HC}	60571	37397	40000
N_{HCG}	42430	2931	40000
Time in HC1	0.5	0.06	0.53

Real Data:

> *We find a very recent speciation time of human-chimp (4.1 $\pm$ 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...*