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

28 May 2013

Scientific Motivation

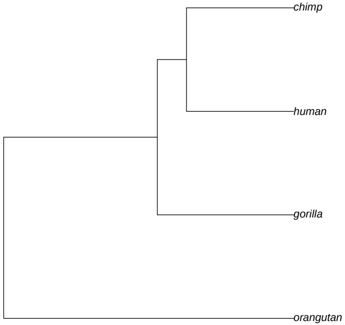
Learn about the evolutionary history of Human, Chimpanzee, Gorilla and Orangutan by comparing their DNA sequences.

```
>human
ACATTTTTGTTTAAATGATACTGACATTTCCTGGGTTGTCCATTTGGAGT...
>chimp
AGATTTTTGTTTAAATGATACTGACATTTCCTGGGTTGTCCATTTGGAGT...
>gorilla
ACATTTTTGTTTGAATGATACTGACATTTCCTGGGTTGTCCATTTGGAGT...
>orangutan
ACATTTTTGTTTAAGTGATACTGACATTTCCTGGGTTGTCCATTTGGATT...
```

Scientific Motivation

Given some aligned DNA sequences, “bread and butter” phylogenetic techniques tell us

- The species tree
- DNA mutation rates
- Estimates of Branch lengths



Scientific Motivation

Want to know finer scale information about the evolutionary history.

- Every locus has a different genealogy
- All these genealogies are consistent with the species tree

Putting them all together would give us the Ancestral Recombination Graph (ARG).

- Every coalescent and recombination event
- Large and complex
- Not Markov when viewed as a process along the DNA sequence

Scientific Motivation

We want to approximate the ARG, with a Coalescent Hidden Markov Model.

- Make a lot of simplifying assumptions
- More information than standard phylogenetic tree, but not as good as the ARG.
- There are other methods, all balance complexity and scalability

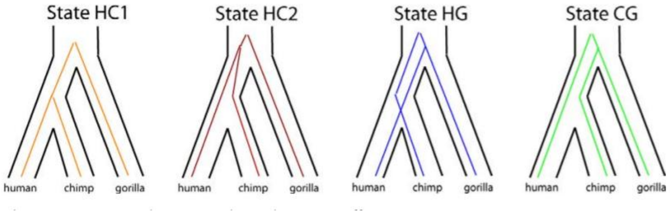
Coal-HMM

All Hidden Markov Models have:

- Hidden States
- Transitions between Hidden States
- Emission Probabilities
- Observed Data

Coal-HMM

- Hidden States
    - There are 4 hidden states.
    - They represent topologies of genealogies found in the ARG



Coal-HMM

- Transition's between Hidden States
  - As we move along the genome, genealogies and hence topological state changes.
  - Transitions are a discrete time Markov Chain with the following probabilities:

$$\mathbf{P} = \begin{bmatrix} 1-3s & s & s & s \\ u & 1-u-2v & v & v \\ u & v & 1-u-2v & v \\ u & v & v & 1-u-2v \end{bmatrix}$$

  - We expect to be in State 1 most of the time

Coal-HMM

- Emission Probabilities
  - What data we see depends on the hidden state: its shape and its branch lengths. $a, b, c, \tilde{a}, \tilde{b}, \tilde{c}$
  - Starting from the common ancestor, DNA mutates down the branches of the tree as a CTMC
  - There are many models for mutation rates, and it probably doesn't matter which one we choose?

Model	df	logLik	AIC	BIC
JC	5.00	-2099529.43	4199068.87	4199129.08
F81	8.00	-2066269.20	4132554.41	4132650.75
SYM	10.00	-2085391.69	4170803.37	4170923.80
GTR	13.00	-2050333.00	4100692.01	4100848.57

Coal-HMM

- Observed Data
  - The aligned DNA sequences
  - The data at some loci are more informative than others

Pattern	Evidence	For State
1100	Strong	1 or 2
1010	Strong	3
0110	Strong	4
110x	Some	1 or 2
101x	Some	3
011x	Some	4
other		None

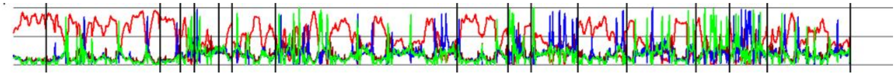
  - 98% of loci are uninformative.
  - 0.007% are strongly informative

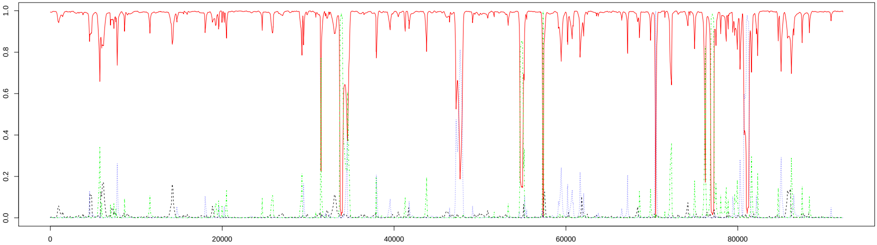
Analysis

1. Set initial parameters
2. Till the likelihood converges:
    1. Propose a likely path through the hidden states, given parameters
    2. Maximum likelihood emission parameters, given the path
    3. Maximum likelihood transition parameters, given the path

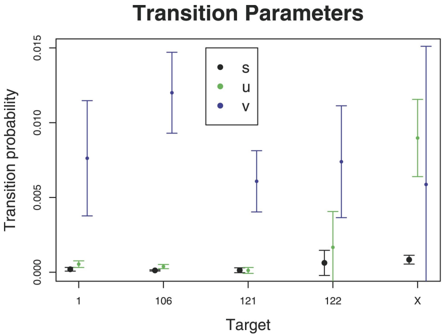
Results

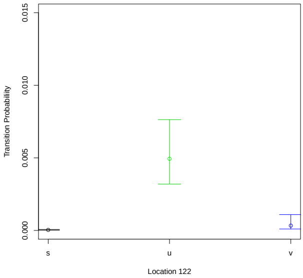
Posterior Probabilities of States



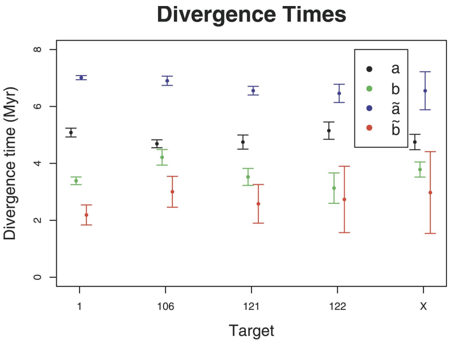


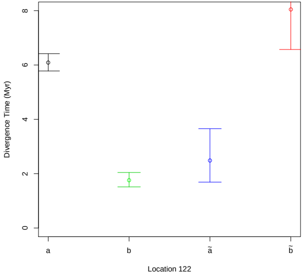
Results


Transition Parameters
Transition probability
0.000
0.005
0.010
0.015
1
106
121
122
X
Target
● s
● u
● v


Transition Probability
0.000
0.005
0.010
0.015
s
u
v
Location 122

Results


Divergence Times
Divergence time (Myr)
0
2
4
6
8
1
106
121
122
X
Target
● a
● b
● ã
● b̃

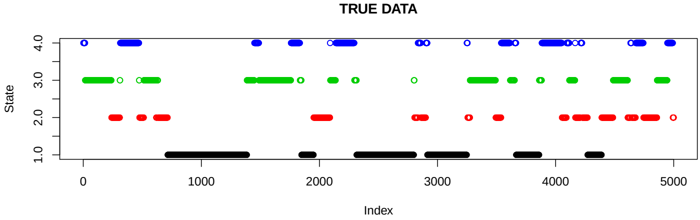

Divergence Time (Myr)
0
2
4
6
8
a
b
ã
b̃
Location 122

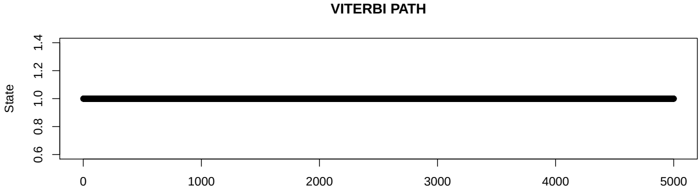
Results

Whats the problem?

The estimates of the transition probabilities are too small.

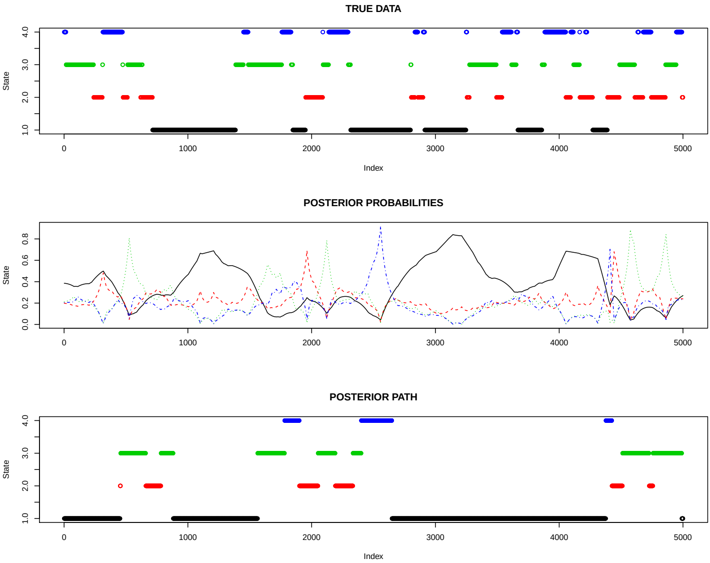
Why?





Results

What about the posterior distribution?


TRUE DATA

POSTERIOR PROBABILITIES

POSTERIOR PATH

Results

- How to solve this problem?
    - More Constraints on the parameters?
    - Try more initial conditions, convergence criteria
- Still To Do:
    - Results for the other three data sets
    - Simulation Study using data simulated from model
    - Simulation Study using data simulated from better model

Conclusions and Criticisms

- Method probably works
  - A lot of the details are left out
- Makes significant simplifying assumptions
  - How much do we trust the conclusions from genetics point of view?
  - Trust methods that are slower, but better approximate the ARG.
- Can be generalised, but...
  - Not easy to add species
  - This will increase model complexity quickly