Markov chain Monte-Carlo to estimate speciation and extinction rates:

making use of the forest hidden behind the (phylogenetic) tree

Nicolas Salamin

Department of Ecology and Evolution
University of Lausanne
Switzerland

Time and rates of divergence

- the temporal dimension contained in a calibrated tree also gives information about the tempo of lineages evolution through the time intervals between splits on the tree
- a formal mathematical description has to be developed to obtain estimates of speciation and extinction rates from a phylogenetic tree
- stochastic models of lineage growth, such as pure birth or birth/death process, have been extensively studied and are ideal for such a task

Investigating speciation rates

Some of the approaches developed so far:

- birth and death process

  e.g. Nee et al. (1994), Bokma (2003)

- survival analysis

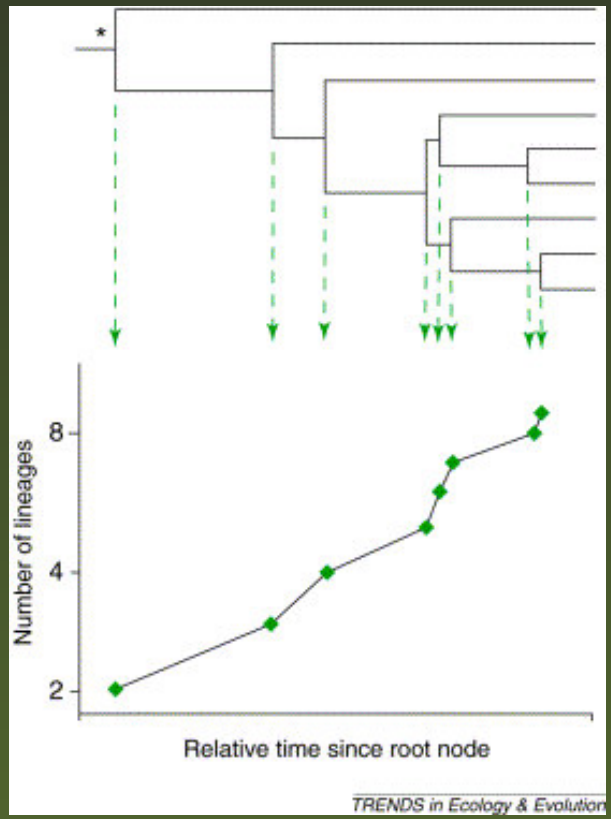
  Paradis (1997, 1998)

- non-parametric approach: distribution of the number of species within clades of a tree suggest lineages with higher diversification rates
  $\longrightarrow$ detection of the morphological or ecological characteristic associated possible

  Slowinski and Guyer (1993), Paradis (2005)

Lineage-through-time plot I

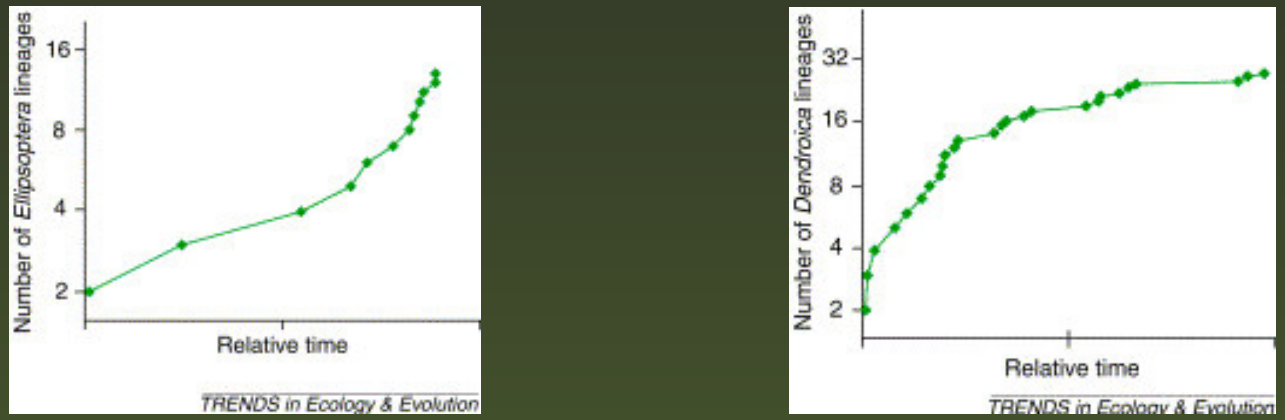
- the most widely used approach is the lineage-through-time plots
  Nee et al. (1994); Nee (2001)



Barraclough and Nee (2001)

Lineage-through-time plot II

- departure from a constant speciation rate



Barraclough and Nee (2001)

- the slope near the present asymptotically approaches the speciation rate, while the slope in the main body of the graph represents only the net speciation rate

Aims

- to develop a full maximum likelihood approach for the estimation of λ, the speciation rate, and μ, the extinction rate
- to develop a framework to take into account the uncertainty about branch length and topology while estimating the speciation and extinction rates
- to assess the method developed through computer simulations

Birth and death process probabilities

- lineages are independent, and rates are constant through time and lineages

- probability that one lineage does not speciate nor get extinct during time t

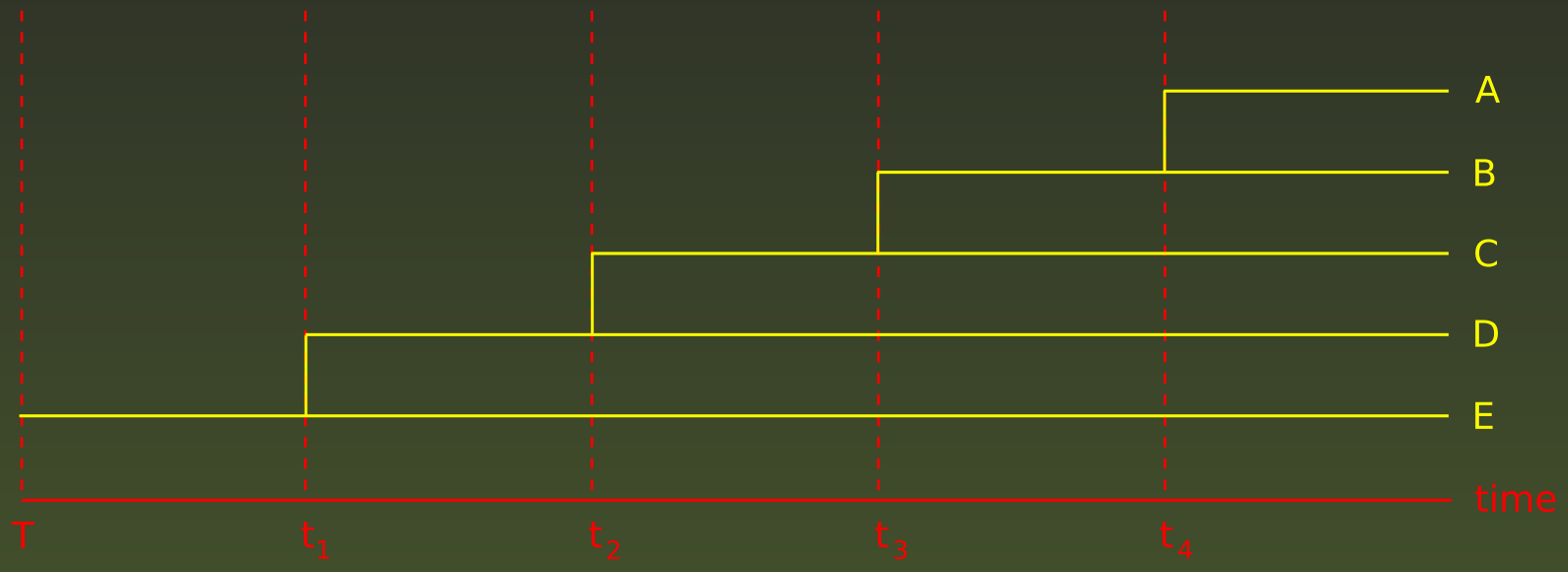
  (Kendall, 1948; Thompson, 1975)

$$p_1(t) = \frac{(\lambda - \mu)^2 e^{(\lambda-\mu)t}}{(\lambda e^{(\lambda-\mu)t} - \mu)^2}$$

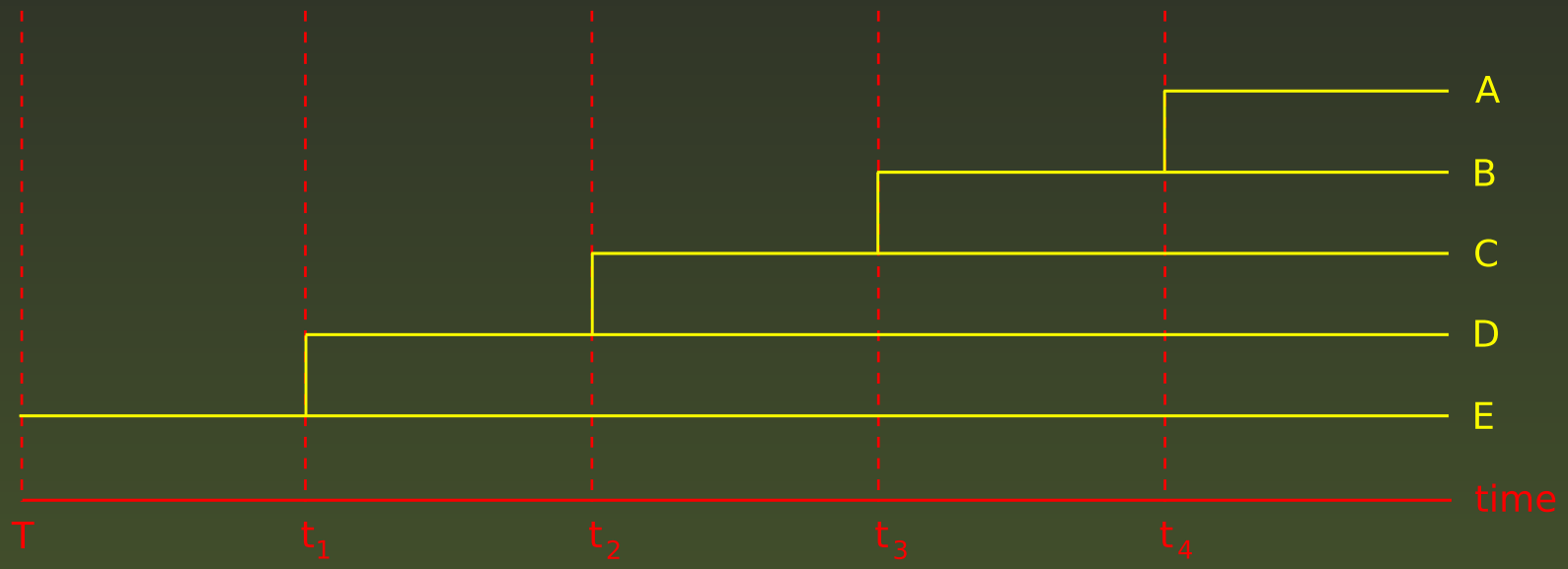
- which, when $\lambda = \mu$, simplify to

$$p_1(t) = \frac{1}{(1 + \lambda t)^2}$$

Probability density of a simple tree



Probability density of a simple tree



- the probability density of this particular tree:

$$Prob(5, t_1, t_2, t_3, t_4 | \lambda, \mu, T) = p_1(T)\lambda dt_1 p_1(t_1)\lambda dt_2 p_1(t_2)\lambda dt_3 p_1(t_3)\lambda dt_4 p_1(t_4)$$

$$= p_1(T)\lambda^4 \prod_{i=1}^{4} p_1(t_i) dt_i$$

Probability density for n lineages

- for any given labelled histories with n species, the probability density becomes:

$$Prob(n, t_1, ..., t_n | \lambda, \mu, T) = p_1(T)\lambda^{n-1}(n-1)! \prod_{i=1}^{n-1} p_1(t_i)dt_i$$

Probability density for n lineages

- for any given labelled histories with n species, the probability density becomes:

$$Prob(n, t_1, ..., t_n | \lambda, \mu, T) = p_1(T) \lambda^{n-1} (n-1)! \prod_{i=1}^{n-1} p_1(t_i) dt_i$$

- by conditioning on having n terminal species, the estimation of $p_1(T)$ can be eluded:

$$Prob(t_1, ..., t_n | n, \lambda, \mu, T) = \prod_{i=1}^{n-1} \frac{p_1(t_i) dt_i}{\int_0^T p_1(u) du}$$

Dealing with phylogenetic uncertainty

- to take into account uncertainty in the branch lengths and the topology, we should sum over all possible trees
- similar to the Bayesian approach (e.g. MrBayes), without specifying a prior distribution on parameters
- similar approach can be used in population genetics to estimate population size, migration rate, recombination rate, selection

  http://evolution.gs.washington.edu/lamarck - Joe Felsenstein's lab

Sampling all possible trees

- we could draw a random sample, but most trees are extremely implausible given the data. Such a random sample would have to be extremely large

Sampling all possible trees

- we could draw a random sample, but most trees are extremely implausible given the data. Such a random sample would have to be extremely large
- we could bootstrap the data, but slow and biased

Sampling all possible trees

- we could draw a random sample, but most trees are extremely implausible given the data. Such a random sample would have to be extremely large
- we could bootstrap the data, but slow and biased
- we can use an importance sampling approach: we concentrate sampling on those trees which are plausible and will contribute substantially to the likelihood of the data and parameters

Sampling all possible trees

- we could draw a random sample, but most trees are extremely implausible given the data. Such a random sample would have to be extremely large
- we could bootstrap the data, but slow and biased
- we can use an importance sampling approach: we concentrate sampling on those trees which are plausible and will contribute substantially to the likelihood of the data and parameters
- very versatile approach to estimate unknown and complex distribution

Monte-Carlo integration on trees

Maximum likelihood estimation of speciation and extinction rate:

- we want to get

$$f(T) = Prob(T|\lambda, \mu) = L(\lambda, \mu)$$

Monte-Carlo integration on trees

Maximum likelihood estimation of speciation and extinction rate:

- we want to get

$$f(T) = Prob(T|\lambda, \mu) = L(\lambda, \mu)$$

- but we can’t

Monte-Carlo integration on trees

Maximum likelihood estimation of speciation and extinction rate:

- we want to get

$$f(T) = Prob(T|\lambda, \mu) = L(\lambda, \mu)$$

- but we can't, so we sample from the posterior probability for some driving values λ_0 and μ_0 of λ and μ

$$g(T) = Prob(T|D, \lambda_0, \mu_0) = \frac{Prob(D|T)Prob(T|\lambda_0, \mu_0)}{Prob(D|\lambda_0, \mu_0)}$$

then,

$$L(\lambda,\mu) = \int_T \frac{f(T)}{g(T)} g(T) dT$$

$$= \int_T \frac{Prob(D|T)}{Prob(D|T)} \frac{Prob(D|\lambda_0,\mu_0)}{Prob(D|\lambda_0,\mu_0)} \frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} L(\lambda_0,\mu_0) dT$$

then,

$$L(\lambda,\mu) = \int_T \frac{f(T)}{g(T)} g(T) dT$$

$$= \int_T \frac{Prob(D|T)}{Prob(D|T)} \frac{Prob(D|\lambda_0,\mu_0)}{Prob(D|\lambda_0,\mu_0)} \frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} L(\lambda_0,\mu_0) dT$$

$$\frac{L(\lambda,\mu)}{L(\lambda_0,\mu_0)} = \int_T \frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} dT = \mathbb{E}_T \left[\frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} \right]$$

then,

$$L(\lambda,\mu) = \int_T \frac{f(T)}{g(T)} g(T) dT$$

$$= \int_T \frac{Prob(D|T)}{Prob(D|T)} \frac{Prob(D|\lambda_0,\mu_0)}{Prob(D|\lambda_0,\mu_0)} \frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} L(\lambda_0,\mu_0) dT$$

$$\frac{L(\lambda,\mu)}{L(\lambda_0,\mu_0)} = \int_T \frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} dT = \mathbb{E}_T \left[\frac{Prob(T|\lambda,\mu)}{Prob(T|\lambda_0,\mu_0)} \right]$$

$$\simeq \frac{1}{n} \sum_{i=1}^{n} \frac{Prob(T_i|\lambda,\mu)}{Prob(T_i|\lambda_0,\mu_0)}$$

which is an average of a likelihood ratio over the n trees T_i sampled.

MCMC: Metropolis-Hastings method

- to draw a sample (or Markov chain) $T_1, ..., T_n$ from a distribution proportional to a function $g(T)$, start from a tree T_0 and then:

MCMC: Metropolis-Hastings method

- to draw a sample (or Markov chain) $T_1, ..., T_n$ from a distribution proportional to a function $g(T)$, start from a tree T_0 and then:
  1. draw a change in T_i from some "proposal distribution": $T_i \mapsto T_{i+1}$

MCMC: Metropolis-Hastings method

- to draw a sample (or Markov chain) $T_1, ..., T_n$ from a distribution proportional to a function $g(T)$, start from a tree T_0 and then:
  1. draw a change in T_i from some "proposal distribution": $T_i \mapsto T_{i+1}$
  2. accept the change if a uniformly-distributed random number r satisfies

$$r < \frac{Prob(T_i|T_{i+1})}{Prob(T_{i+1}|T_i)} \frac{L(D|T_{i+1})Prob(T_{i+1}|\lambda_0, \mu_0)}{L(D|T_i)Prob(T_i|\lambda_0, \mu_0)}$$

MCMC: Metropolis-Hastings method

- to draw a sample (or Markov chain) $T_1, ..., T_n$ from a distribution proportional to a function $g(T)$, start from a tree T_0 and then:
  1. draw a change in T_i from some "proposal distribution": $T_i \mapsto T_{i+1}$
  2. accept the change if a uniformly-distributed random number r satisfies

$$r < \frac{Prob(T_i|T_{i+1})}{Prob(T_{i+1}|T_i)} \frac{L(D|T_{i+1})Prob(T_{i+1}|\lambda_0, \mu_0)}{L(D|T_i)Prob(T_i|\lambda_0, \mu_0)}$$

  3. go back to 1. many times (e.g. 1 mio times or ‘generations’)

MCMC: Metropolis-Hastings method

- to draw a sample (or Markov chain) $T_1, ..., T_n$ from a distribution proportional to a function $g(T)$, start from a tree T_0 and then:
  1. draw a change in T_i from some "proposal distribution": $T_i \mapsto T_{i+1}$
  2. accept the change if a uniformly-distributed random number r satisfies

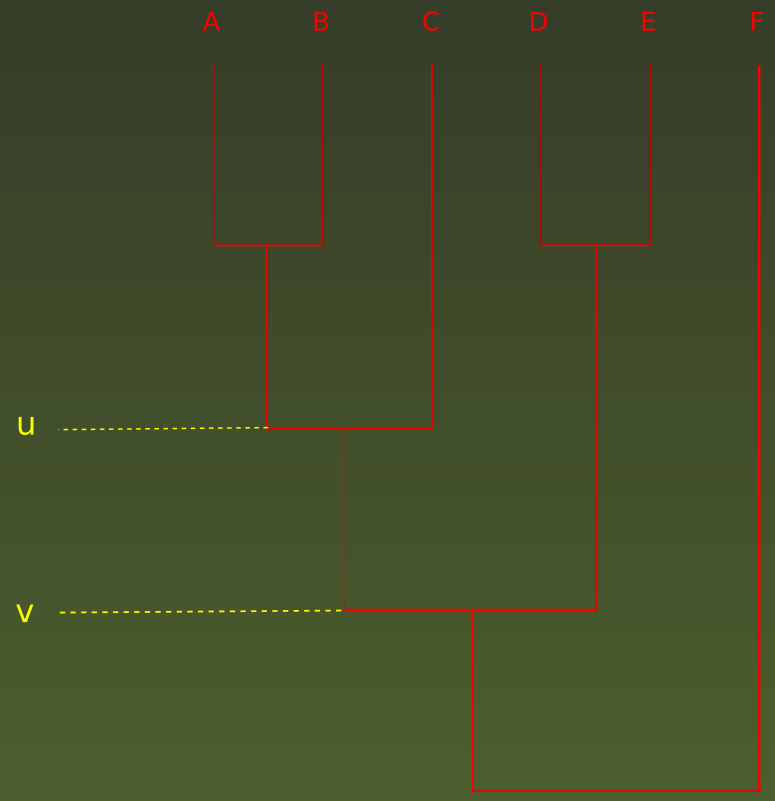
$$r < \frac{Prob(T_i|T_{i+1})}{Prob(T_{i+1}|T_i)} \frac{L(D|T_{i+1})Prob(T_{i+1}|\lambda_0, \mu_0)}{L(D|T_i)Prob(T_i|\lambda_0, \mu_0)}$$

  3. go back to 1. many times (e.g. 1 mio times or ‘generations’)

- If we do this long enough and other nice statistical properties hold, the set of T_i will be a sample from the right distribution

The "proposal distribution"

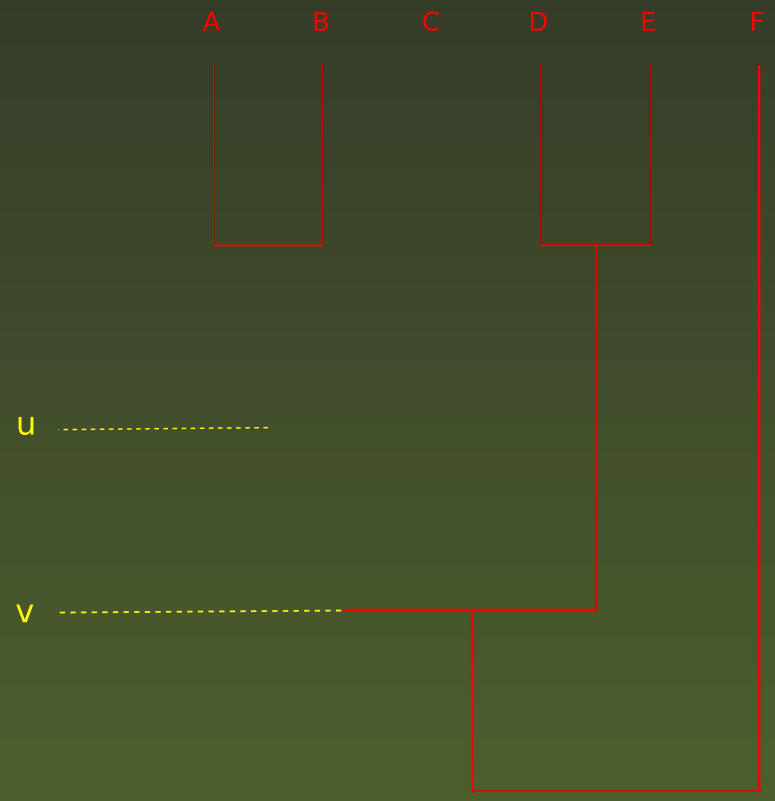
- the idea is to wander through the tree space and calculate for each change the probability of going from one tree to another:



- pick a node u and its ancestor v

The "proposal distribution"

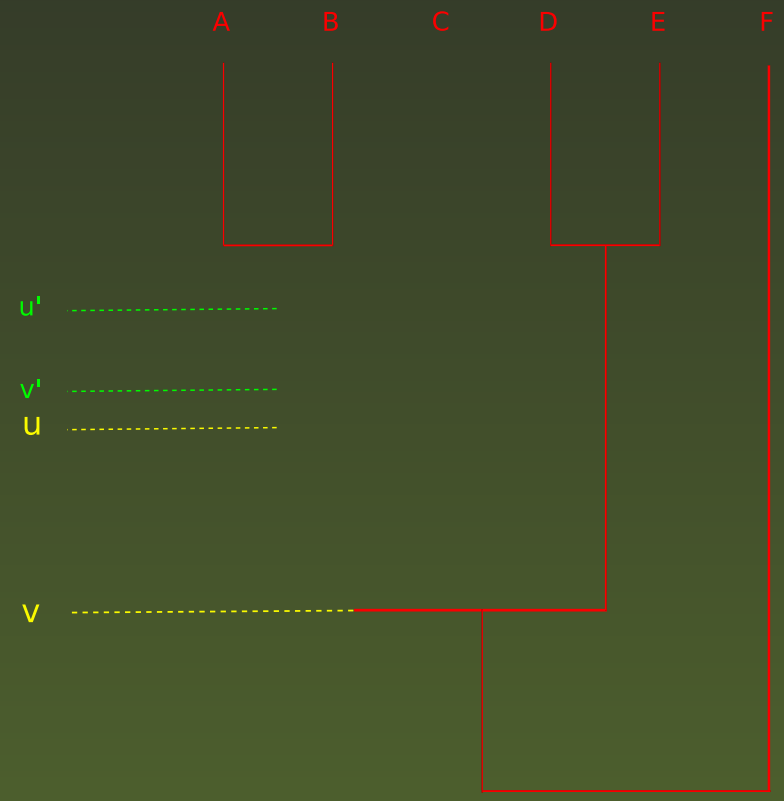
- the idea is to wander through the tree space and calculate for each change the probability of going from one tree to another:



- erase the branches connected to u

The "proposal distribution"

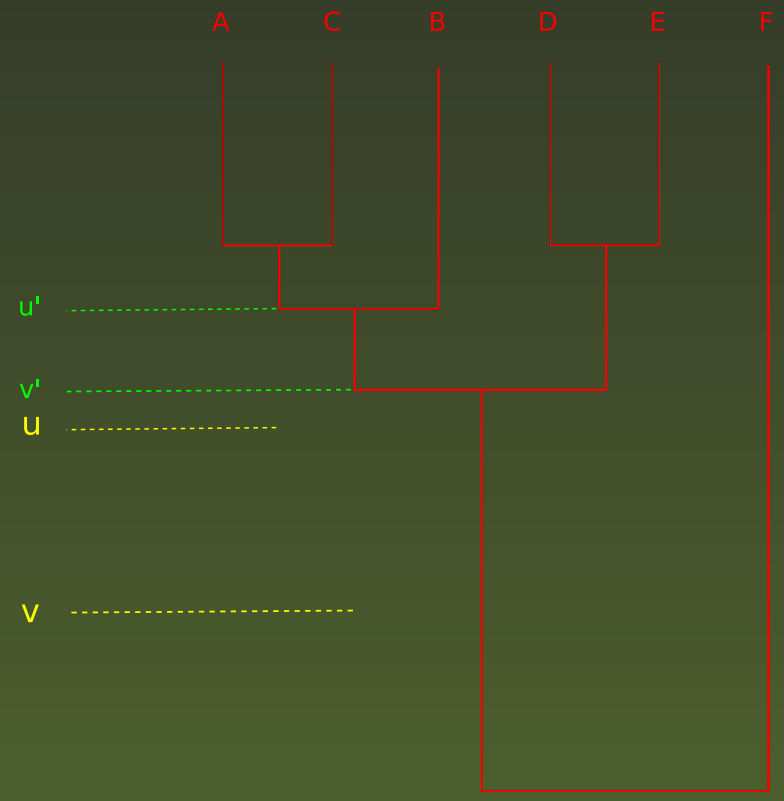
- the idea is to wander through the tree space and calculate for each change the probability of going from one tree to another:



- draw new heights u' and v'

The "proposal distribution"

- the idea is to wander through the tree space and calculate for each change the probability of going from one tree to another:



- reattach the nodes

Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis

Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis
- the solution taken so far is to update the driving values through multiple successive chains:
  - start with some values of λ_0 and μ_0

Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis
- the solution taken so far is to update the driving values through multiple successive chains:
  - start with some values of λ_0 and μ_0
  - run 10 short chains (e.g. 100,000 generations), updating the values of λ and μ after each chain

Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis
- the solution taken so far is to update the driving values through multiple successive chains:
  - start with some values of λ_0 and μ_0
  - run 10 short chains (e.g. 100,000 generations), updating the values of λ and μ after each chain
  - run 1 long chain (e.g. 10,000,000 generations), to update λ and μ one last time

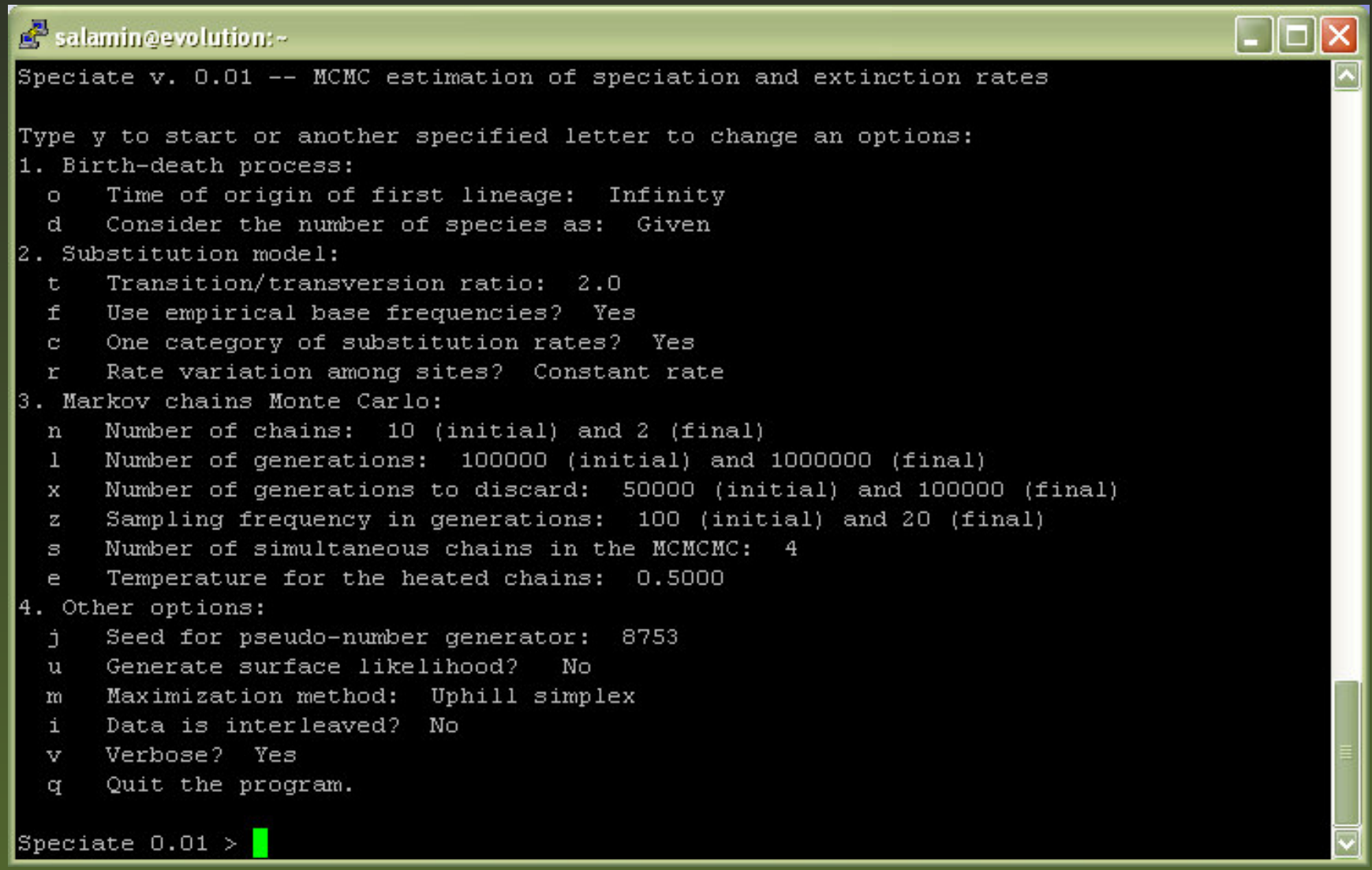
Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis
- the solution taken so far is to update the driving values through multiple successive chains:
  - start with some values of λ_0 and μ_0
  - run 10 short chains (e.g. 100,000 generations), updating the values of λ and μ after each chain
  - run 1 long chain (e.g. 10,000,000 generations), to update λ and μ one last time
  - run 1 long chain (e.g. 10,000,000 generations), and use the trees sampled as the distribution of interest

Getting the right driving values

- ideally driving values used to sample the trees T_i should be the (unknown) MLE of λ and μ
- if not, a bias is introduced in the analysis
- the solution taken so far is to update the driving values through multiple successive chains:
  - start with some values of λ_0 and μ_0
  - run 10 short chains (e.g. 100,000 generations), updating the values of λ and μ after each chain
  - run 1 long chain (e.g. 10,000,000 generations), to update λ and μ one last time
  - run 1 long chain (e.g. 10,000,000 generations), and use the trees sampled as the distribution of interest
  - maximize the likelihood surface to get the estimated values of λ and μ and their confidence interval

Speciate – http://www2.unil.ch/phylo



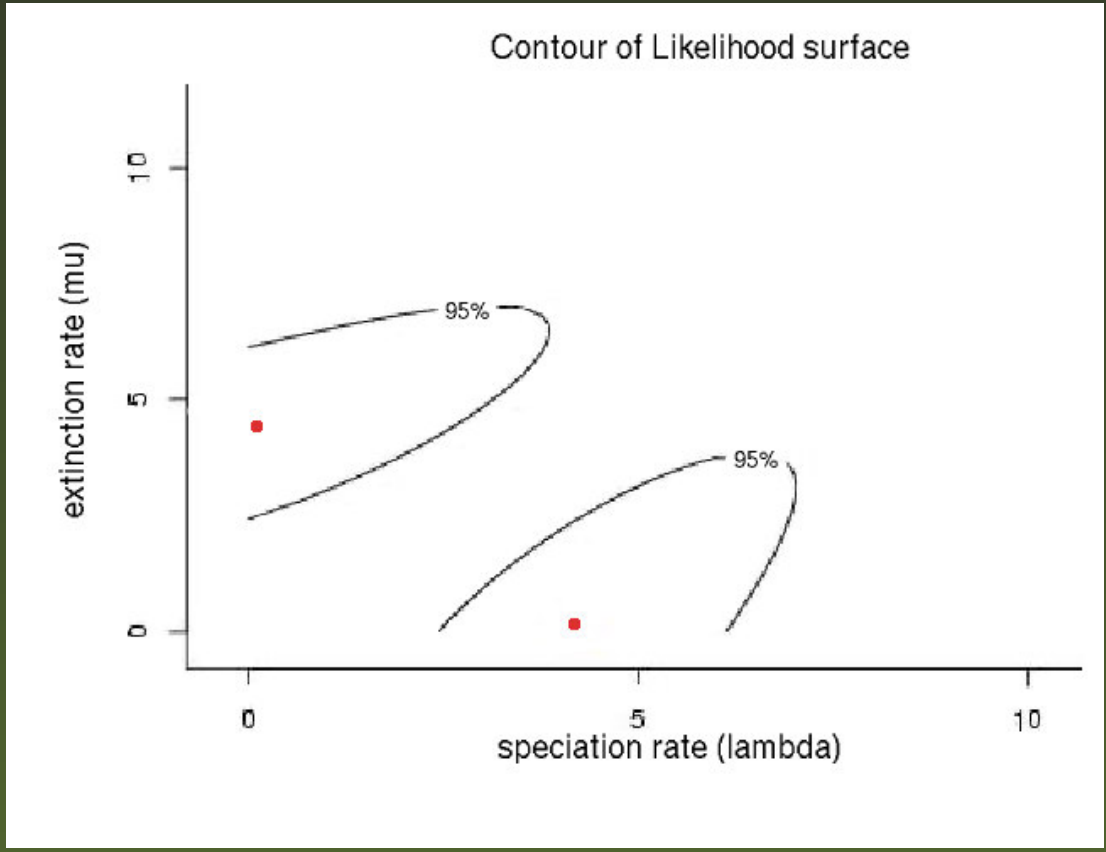
Simulations

Assessment of the current method through computer simulation:

- define the intervals of time between speciation events using a birth/death process with parameters λ and μ in expected number of substitution
- based on the times obtained, create corresponding topologies
- simulate DNA sequences, by evolving the different sites on the trees according to the GTR+Γ model of evolution
- for each set of DNA sequences, estimate the values of $\hat{\lambda}$ and $\hat{\mu}$ and their confidence intervals

20 species, 1 replicate

- 1000 nucleotides, F84+Γ model used, $\lambda = 1.200$, $\mu = 0.800$, estimated $\hat{\lambda} = 4.645$ and $\hat{\mu} = 0.034$



20 species, 100 replicates

- 1000 nucleotides simulated, F84+Γ model used

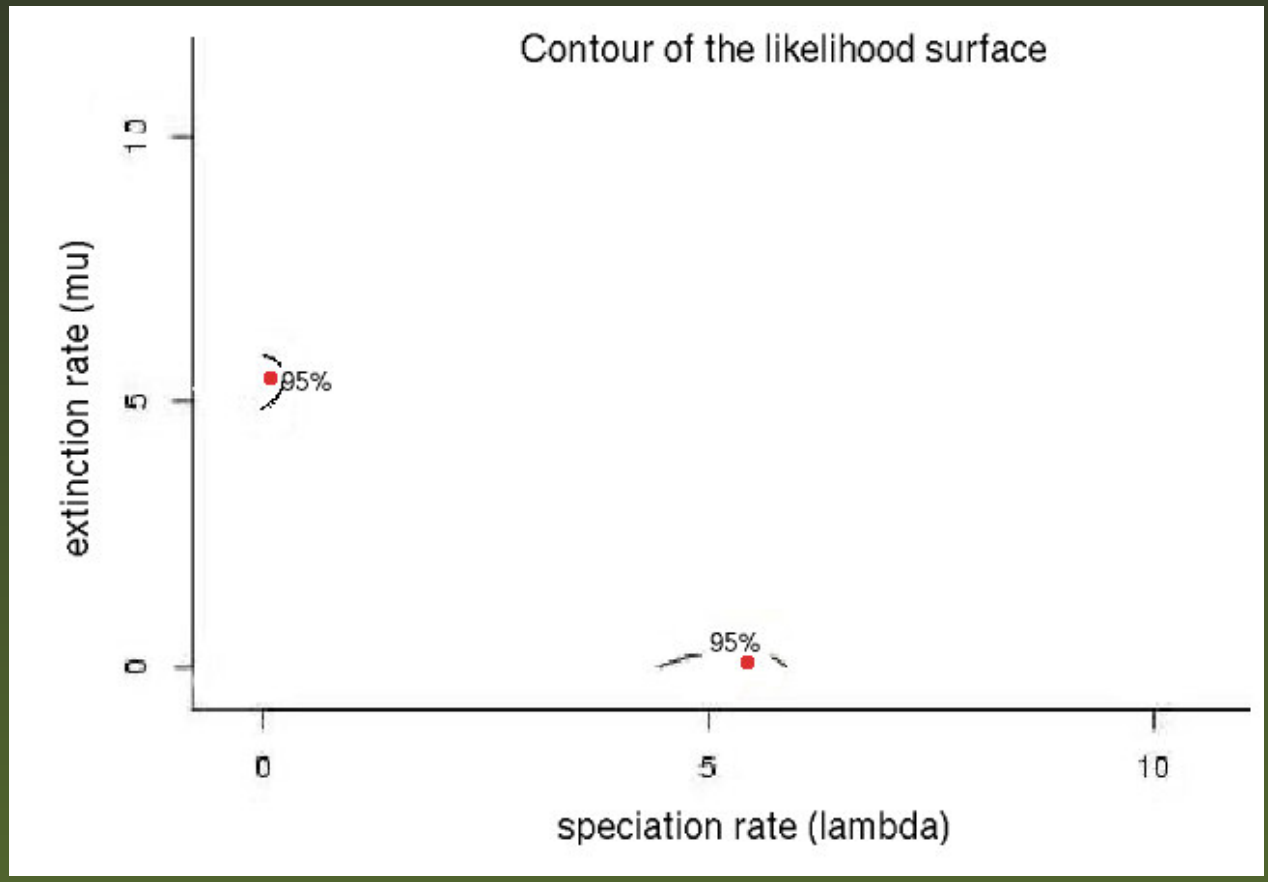
True λ	True μ	Mean $\hat{\lambda}$	Mean $\hat{\mu}$	SD $\hat{\lambda}$	SD $\hat{\mu}$
0.800	0.800	0.343	0.001	1.342	1.567
1.200	0.800	2.345	0.004	2.345	1.342
5.000	0.020	8.453	0.000	3.245	1.597

- 10000 nucleotides simulated, F84+Γ model used

True λ	True μ	Mean $\hat{\lambda}$	Mean $\hat{\mu}$	SD $\hat{\lambda}$	SD $\hat{\mu}$
0.800	0.800	0.566	0.236	0.846	1.084
1.200	0.800	1.543	0.334	0.453	0.984
5.000	0.020	4.576	1.324	1.087	1.234

200 species, 1 replicate

- 1000 nucleotides, F84+Γ model used, $\lambda = 5.000$, $\mu = 0.020$, estimated $\hat{\lambda} = 5.156$ and $\hat{\mu} = 0.067$



200 species, 100 replicates

- 1000 nucleotides simulated, F84+Γ model used

True λ	True μ	Mean $\hat{\lambda}$	Mean $\hat{\mu}$	SD $\hat{\lambda}$	SD $\hat{\mu}$
0.800	0.800	1.301	1.023	0.345	0.456
1.200	0.800	1.498	0.764	0.234	0.457
5.000	0.020	3.245	0.000	1.234	0.324

- 10000 nucleotides simulated, F84+Γ model used

True λ	True μ	Mean $\hat{\lambda}$	Mean $\hat{\mu}$	SD $\hat{\lambda}$	SD $\hat{\mu}$
0.800	0.800	0.904	0.645	0.123	0.213
1.200	0.800	1.023	0.986	0.046	0.127
5.000	0.020	4.675	0.295	0.233	0.345

Conclusions

- conditioning on the number of species gives better parameter estimates
  $\rightarrow$ remove the problem of obtaining the time of origin of the first lineage; although information is lost
- it becomes impossible to distinguish cases where $\lambda > \mu$ or $\mu > \lambda$
- large amount of nucleotides are required to obtain accurate estimates
- many terminal species are also required in order to have enough information to estimate the parameters of the stochastic process

Some considerations

- strong assumption of a molecular clock required

- how to chose the driving values? Possible other solutions to investigate
  - sampling from a mixture distribution
  - stochastic approximation before starting the chain
  - bridge sampling

Further work

- implementing more complex model, e.g. quantitative model of evolution of the rate of speciation and extinction

Further work

- implementing more complex model, e.g. quantitative model of evolution of the rate of speciation and extinction

- model selection
  - can’t rely on likelihood ratio test
  - reversible jump Monte Carlo

Further work

- implementing more complex model, e.g. quantitative model of evolution of the rate of speciation and extinction
- model selection
  - can’t rely on likelihood ratio test
  - reversible jump Monte Carlo
- Further MCMC approach: sampling the tree space and more...
  - incomplete sampling of lineages
  - Hypothesis testing by combining rates of speciation with morphological, ecological changes

Acknowledgements

- Joe Felsenstein

  Genome Science Department, University of Washington, Seattle, USA
- Elizabeth Thompson

  Statistics Department, University of Washington, Seattle, USA
- Mary Kuhner, Jon Yamato and Chul-Joo Kang

  Genome Science Department, University of Washington, Seattle, USA
- Bruce Rannala

  Medical Genetics Department, University of Alberta, Edmonton, Canada
- Brendan Murphy

  Statistics Department, Trinity College, Dublin, Ireland

Funding:

- Swiss National Science Foundation

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION