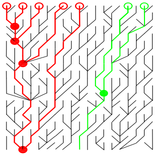
Evolution of the rate of evolution

—

An analytical solution to the compound Poisson process



Stéphane Guindon

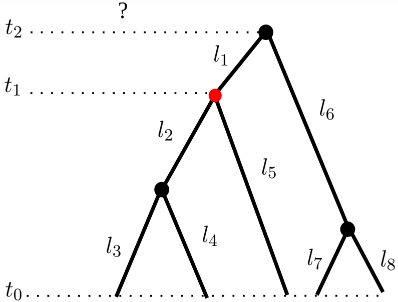
Department of Statistics, University of Auckland, New Zealand.
LIRMM, UMR 5506 CNRS Montpellier, France.

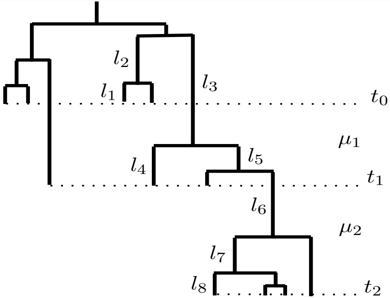
Outline

A bit of history...

- Linus Pauling and Emile Zuckerkandl (1962): "*molecular clock hypothesis*".
- Allan Wilson (1967): molecular dating under the molecular clock assumption.
- 30 years passed...
- Michael Sanderson (1997) and Jeffrey Thorne (1998): estimation of evolutionary divergence times without the restriction of a uniform rate across lineages.

Molecular clock rate and time estimation





$$l_5 = \mu \times (t_1 - t_0)$$

$$\hookrightarrow \mu = \frac{l_5}{t_1 - t_0}$$

$$t_2 = \frac{l_1 + l_2 + l_3}{\mu} + t_0$$

$$\mu_1 = \frac{l_4 + l_3 - l_1 - l_2}{t_0 - t_1}$$

$$\mu_2 = \frac{l_5 + l_6 + l_7 + l_8 - l_4}{t_1 - t_2}$$

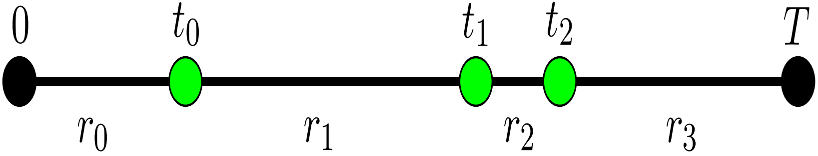
Beyond the molecular clock

- *Local clocks*
  - Substitution rate is organised into a small number of classes,
  - Assign each branch to one of these classes.
- *Penalized likelihood*
  - $\Psi(R,T)$: penalty term for rate changes,
  - Maximise $log(P(D|R,T)) - \lambda\Psi(R,T)$.
- *Bayesian approaches*
  - Explicit stochastic models of the evolution of the substitution rate.
  - Rate trajectory is continuous or discrete.

Models of rate evolution (1/2)

- Log-normal model
  - μ is the mean of the rate at the nodes that begin and end the branch ($r(0)$ and $r(T)$).
  - $log(r(T)) \sim \mathcal{N}(log(r(0)), \nu T)$.
  - Logarithm of the rate undergoes *Brownian motion*.
  - Correlation of mean rates on adjacent branches.
- Exponential model
  - $\mu \sim Exp(\phi)$.
  - No correlation of mean rates.
  - Shape of the distribution does not depend on time duration.

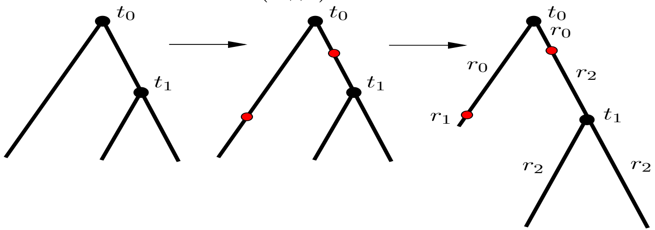
Models of rate evolution (2/2)



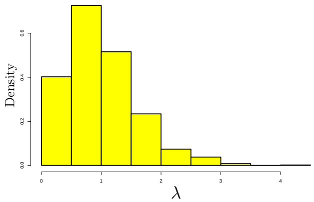
- Compound Poisson process
  - Rates change in discrete jumps.
  - $r(t) \sim \Gamma(\alpha, \beta)$
  - Number of jumps: $n(T) \sim Poisson(\lambda T)$
  - Correlation of mean rates across branches: governed by λ.
  - λT large: distribution of mean rate is approximately Normal.

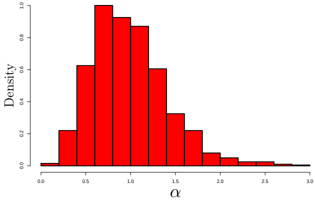
Implementation of the compound Poisson process

- "Jump" event: $Poisson(\lambda \Delta t)$
- Substitution rates: $\Gamma(\alpha, \beta)$



- MCMC → posterior distribution of λ and α



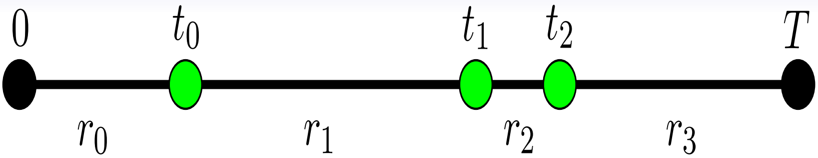


Advantages and drawbacks

- Log-normal
  - Computationally tractable
  - Crude (deterministic) description of the mean rates.
  - Biologically relevant ?
- Exponential
  - Computationally tractable.
  - Distribution of mean substitution rate does not depend on time duration.
  - No correlation of mean rates across branches.
- Compound Poisson
  - Description of rate changes plausible from a biological perspective.
  - Elegant way to account for correlation of mean rates across branches.
  - No analytical solution.

Outline

First question

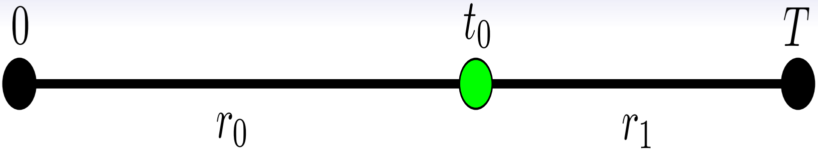


- $r_i \sim \Gamma(\alpha, \beta)$. Hence, $E(r_i) = \alpha\beta$, $V(r_i) = \alpha\beta^2$.
- $n \sim Poisson(\lambda T)$.
- $\mu = \sum_{i=0}^{n} k_i r_i$, where $k_i = \frac{\Delta t_i}{T}$.

What is the distribution of μ ?

- Work out the distribution of μ for a given value of n.
- $\mu = \sum_{i=0}^{n} k_i r_i$ is well approximated by a Gamma distribution.

One jump

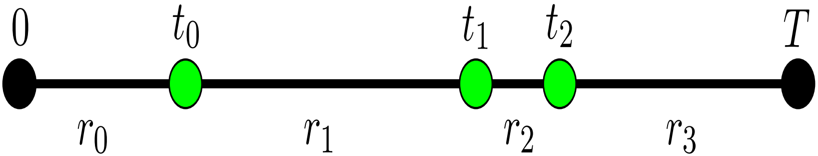


- $\mu = k_0 r_0 + (1 - k_0) r_1$
- Distribution of $t_0 = k_0 T$?

$$P(t_0 = x | n = 1) = \frac{\lambda e^{-\lambda x} \times e^{-\lambda(T-x)}}{\lambda T e^{-\lambda T}} = \frac{1}{T}.$$

- $k_0 \sim U[0,1] \rightarrow E(k_0) = \frac{1}{2}$ and $V(k_0) = \frac{1}{12}$.
- $E(\mu) = E(k_0)E(r_0) + E(1 - k_0)E(r_1) = \alpha\beta$.
- $V(\mu) = V(k_0 r_0) + V((1 - k_0) r_1) + 2Cov(k_0 r_0, (1 - k_0) r_1) = \frac{2}{3}\alpha\beta^2$.

$n \geq 1$ jumps

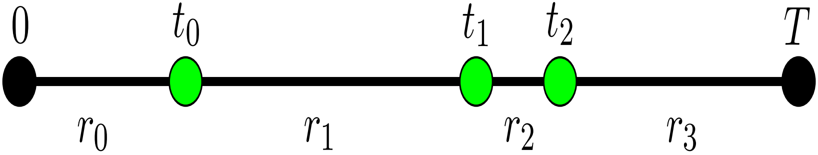


- Distribution of k_0 ?

$$P(t_0 = x | n = y) = \frac{\lambda e^{-\lambda x} \times (\lambda(T-x))^{y-1} e^{-\lambda(T-x)}}{(\lambda T)^y e^{-\lambda T}/y!}$$
$$= \frac{y}{T^y}(T-x)^{y-1}.$$

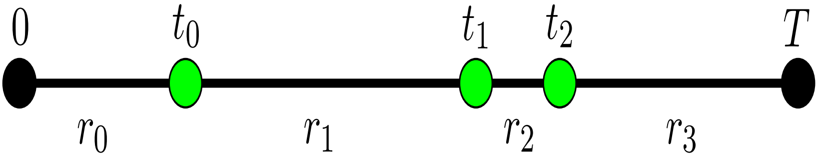
- After little algebra...
  - $E(k_0) = \frac{1}{n+1}$,
  - $E(k_0^2) = \frac{2}{(n+1)(n+2)}$.

$n \geq 1$ jumps



- $\mu = k_0 r_0 + k_1 r_1 + k_2 r_2 + k_3 r_3$.
- $\mu_n = k_0 r_0 + (1 - k_0)\mu_{n-1}$.
- $E(\mu_n) = E(k_0)E(r_0) + E(1 - k_0)E(\mu_{n-1}) \rightarrow \boxed{E(\mu_n) = \alpha\beta}$.

$n \geq 1$ jumps



- The variance is a bit more challenging but can be done.

$$V(\mu_n) = \frac{2\alpha\beta^2 + n(n+1)V(\mu_{n-1})}{(n+1)(n+2)}$$

- Solve the recursion:

$$\boxed{V(\mu_n) = \frac{2}{n+2}\alpha\beta^2}$$

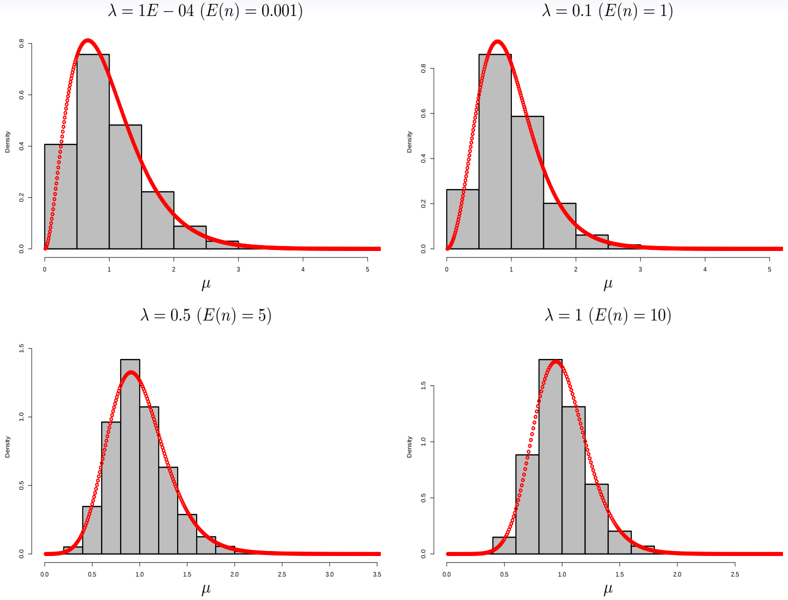
Likelihood calculation

- Data:
  - l, an expected number of substitutions.
  - T, elapsed time.
- $\mu = l/T$
- Likelihood:

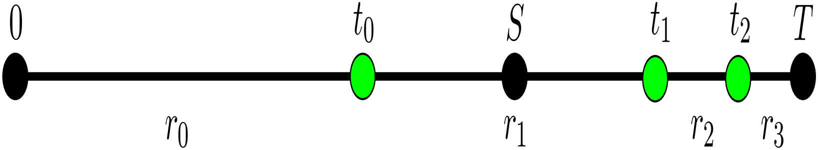
$$p_\mu(u|\lambda, \alpha, \beta, T) = \sum_{n=0}^{\infty} P(n|\lambda, T) p_{\mu_n}(u|\alpha, \beta, n)$$

- $P(n|\lambda, T)$: Poisson distribution with mean and variance λT.
- $p_{\mu_n}(u|\alpha, \beta, n)$: Gamma distribution with mean $\alpha\beta$, and variance $\frac{2}{n+2}\alpha\beta^2$.

The approximation seems good



Second question

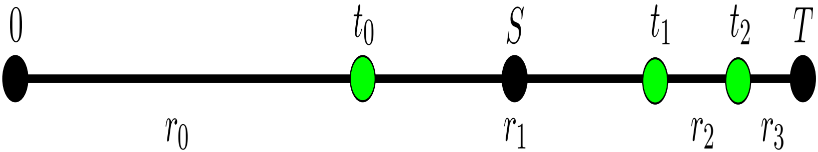


- Two adjacent time intervals: $[0, S]$ and $[S, T]$.
- μ_1 and μ_2 mean rates in $[0, S]$ and $[S, T]$ respectively.
- μ_1 and μ_2 are correlated because of r_1.

What is the joint distribution of μ_1 and μ_2 ?

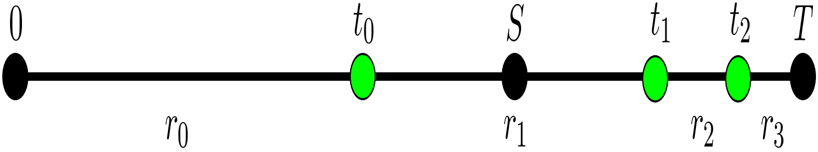
- Work out the density $p_{\mu_2|\mu_1}(u_2|u_1, \lambda, \alpha, T - S)$.

Second question



- I was unable to find an analytical expression...
- First idea: integrate over t_0 in $[0, S]$, t_1 in $[S, T]$ and r_1 in $[0, \infty]$...
- ...didn't work.
- Second idea: use an approximation.
  - 'Many' jumps in $[0, T]$: μ_1 and μ_2 are independent.
  - No jump in $[0, T]$: $p_{\mu_2|\mu_1}(u_2|u_1) = 1$ if $u_2 = u_1$.

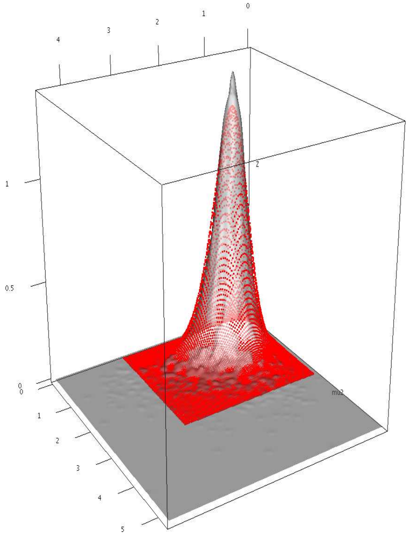
Second question



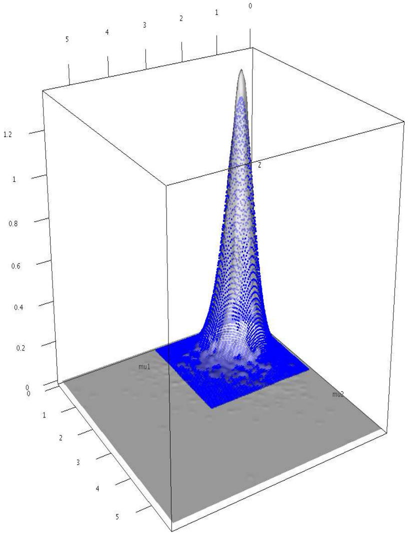
- Use a *mixture model*:
  - $\mu_2|\mu_1 \sim \mathcal{N}(\mu_1, 0.01)$ with probability $P(n = 0|\lambda, T)$,
  - $\mu_2|\mu_1 \sim \mathcal{N}(\mu_1, 0.04)$ with probability $P(n = 1|\lambda, T)$,
  - $\mu_2|\mu_1 \sim \mathcal{N}(\mu_1, 0.09)$ with probability $P(n = 2|\lambda, T)$,
  - μ_2 independent from μ_1 with probability $P(n > 2|\lambda, T)$.

$p_{\mu_1,\mu_2}(u_1, u_2|\lambda, \alpha, T)$, $E(n) = 10$

Mixture

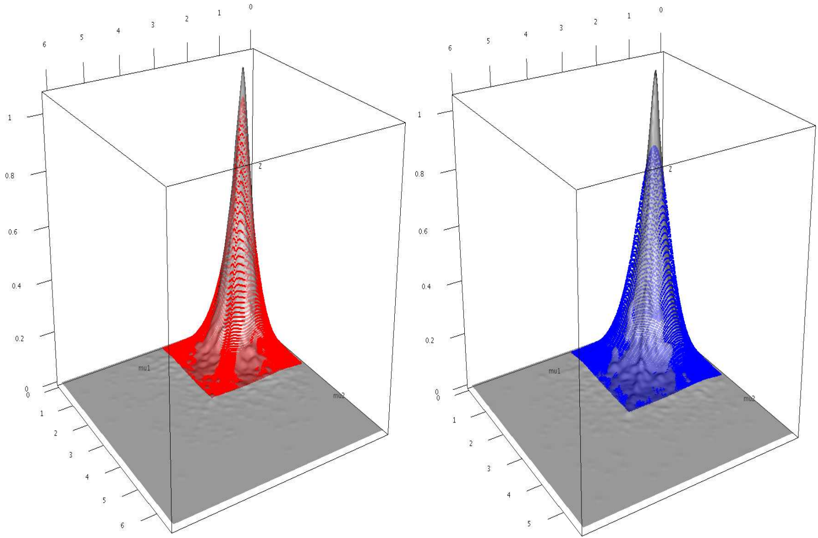


Independent



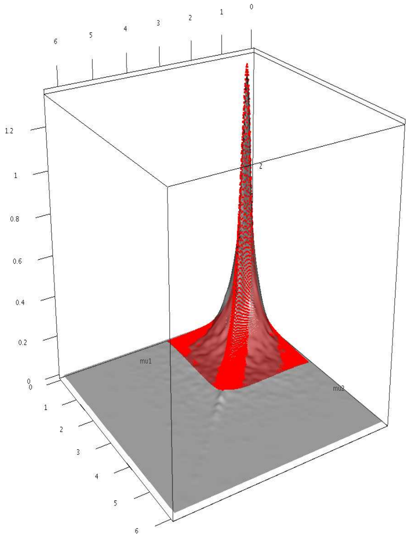
$p_{\mu_1,\mu_2}(u_1, u_2|\lambda, \alpha, T)$, $E(n) = 4$

Mixture Independent

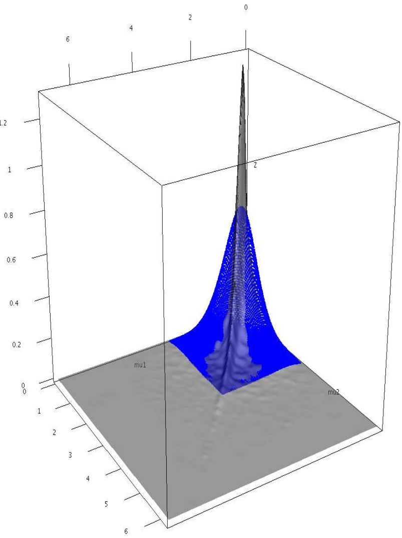


$p_{\mu_1,\mu_2}(u_1, u_2|\lambda, \alpha, T)$, $E(n) = 2$

Mixture

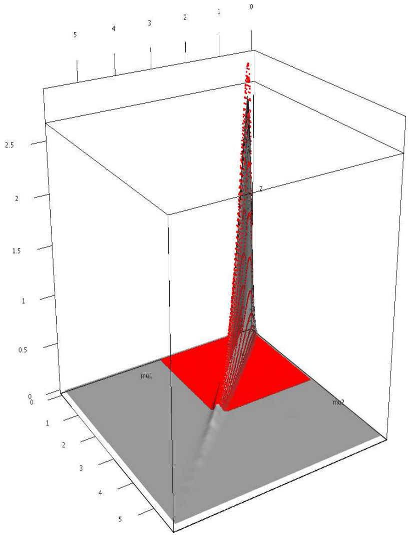


Independent

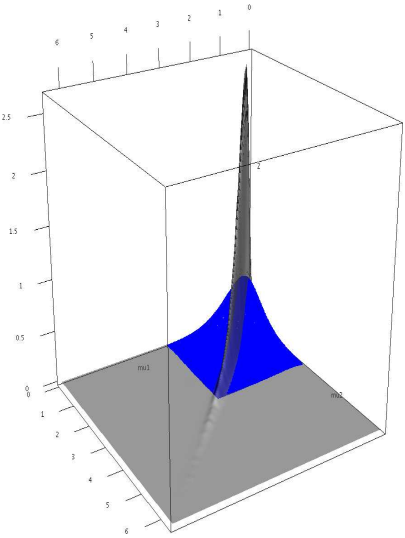


$p_{\mu_1,\mu_2}(u_1, u_2|\lambda, \alpha, T)$, $E(n) = 0.002$

Mixture



Independent



Acknowledgements

- The University of Auckland.
- Dumont d'Urville programme:
  - Ministry of Research, Science & Technology, New Zealand.
  - EGIDE, France.
- Allen Rodrigo, Olivier Gascuel and Vincent Lefort.