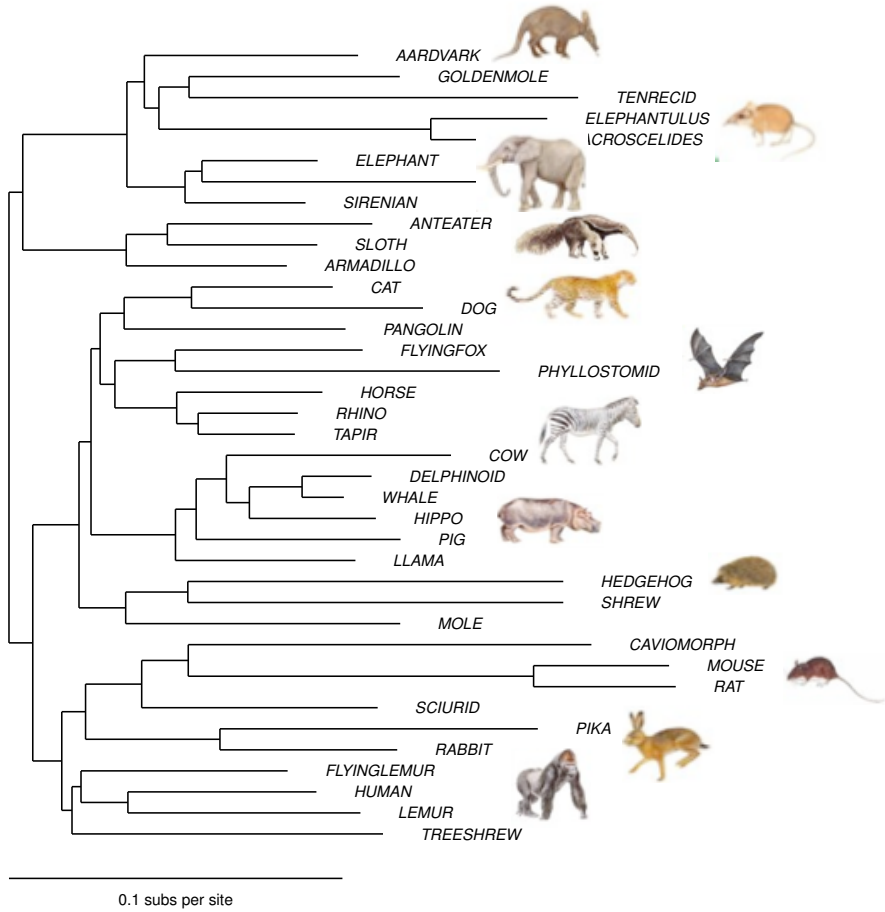
The comparative method; models of continuous trait evolution

Nicolas Lartillot

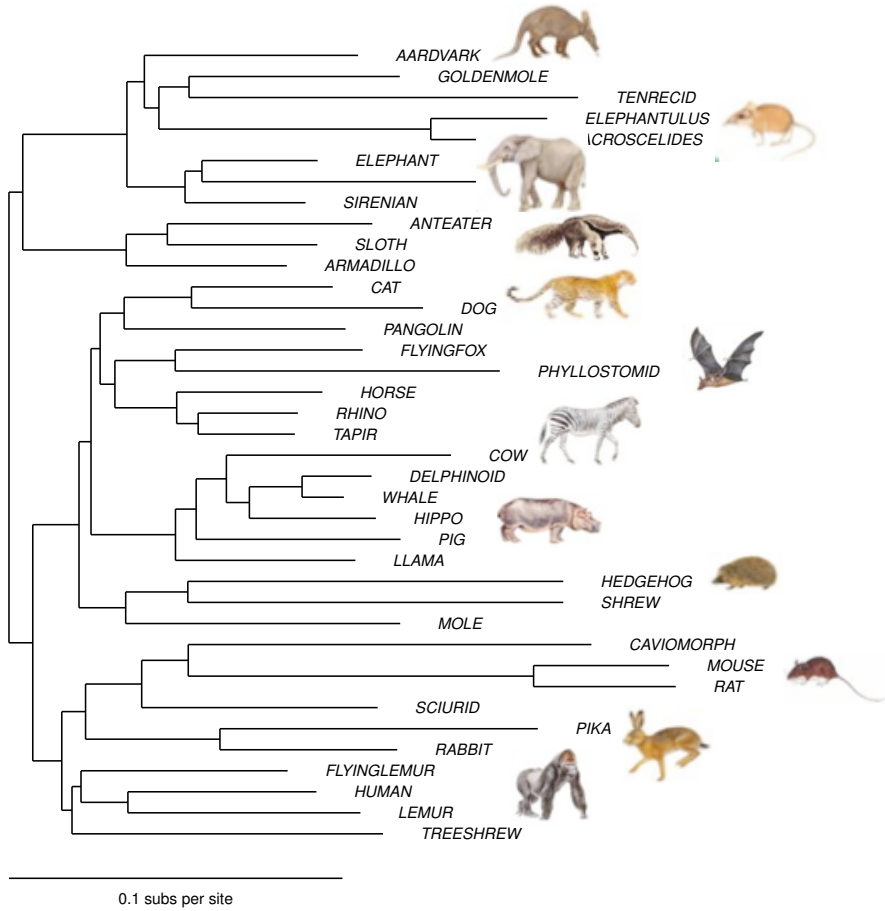
August 29, 2014

Trait evolution on a macroevolutionary scale



- does body size evolve at a constant rate?
- does body size fluctuate around some optimum (or optima)?
- are there bursts, jumps or trends?
- coupled with diversification rates? with substitution rate?

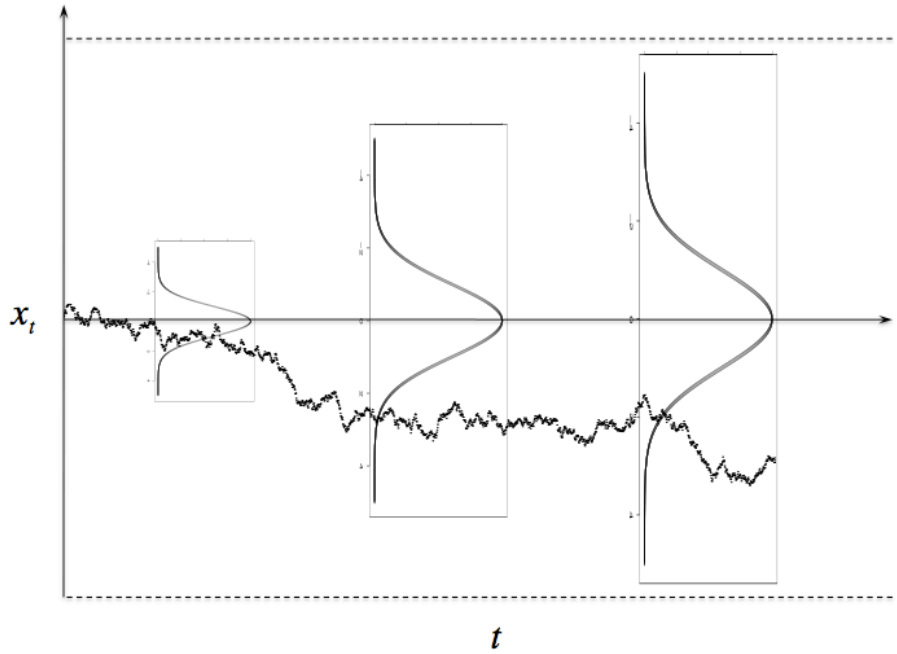
Trait evolution on a macroevolutionary scale



The comparative method

- a model-based, process-oriented approach to these questions
- define a process, or alternative processes, of trait evolution
- condition on data, estimate parameters, test and compare models

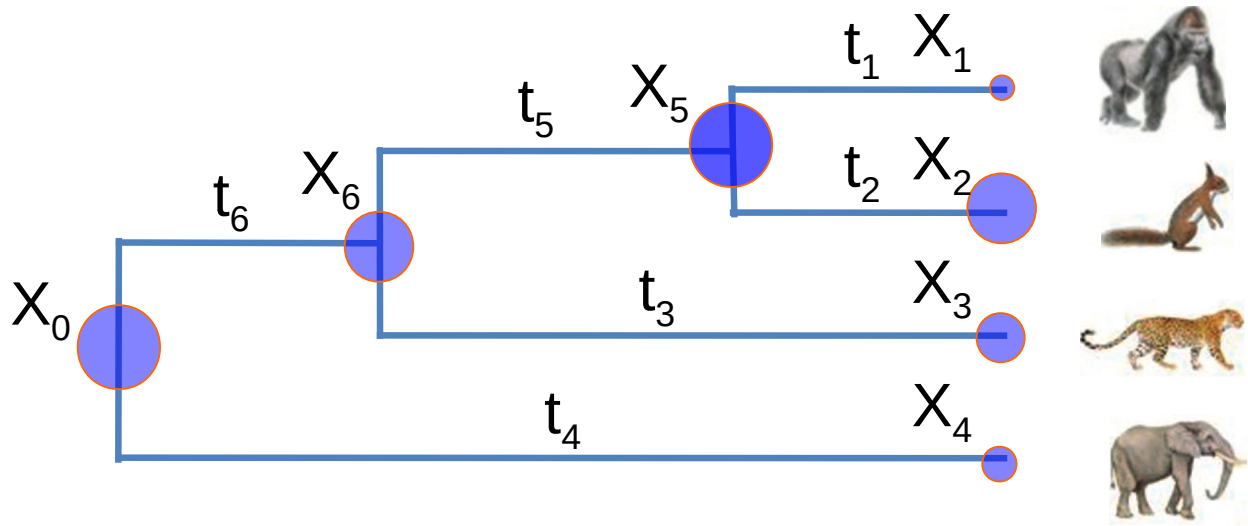
Brownian motion



$$X_t \sim Normal(X_0, \sigma^2 t)$$

- scaling limit of random walk with independent increments
- (or zero mean and finite variance)
- Brownian motion is a Markov process

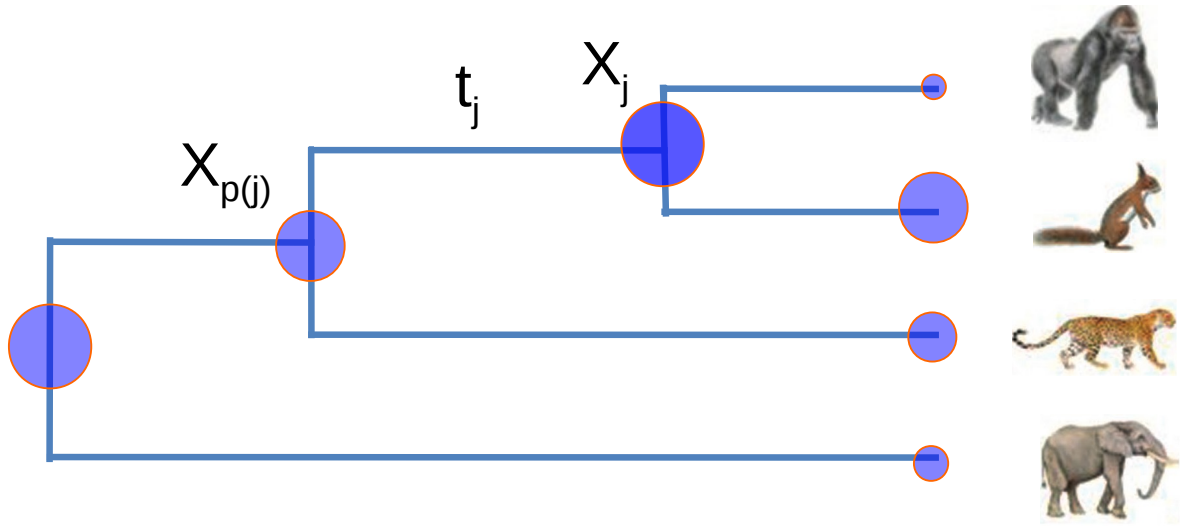
Markov process of trait evolution



Notations and indexing convention

- phylogeny Ψ, n taxa ($2n-1$ nodes, $2n-2$ branches)
- 0: root
- $[1..n]$: tips
- $[n+1..2n-2]$: interior nodes
- X_j: trait value at node j
- t_j: length (in geological time) of branch leading to node j
- $p(j)$: index of parent of node j

Brownian model of trait evolution



Markov decomposition

$$p(X_{1..2n-2} \mid X_0, \Psi, \sigma) = \prod_{j=1}^{2n-2} p\left(X_j \mid X_{p(j)}, t_j, \sigma\right)$$

$$X_j \mid X_{p(j)}, t_j, \sigma \sim \text{Normal}\left(X_{p(j)}, \sigma^2 t_j\right)$$

Uninformative priors: "objective Bayes"

Likelihood

$$p(X_{1..2n-2} \mid X_0, \Psi, \sigma) \quad = \quad \prod_{j=1}^{2n-2} p\left(X_j \mid X_{p(j)}, t_j, \sigma\right)$$

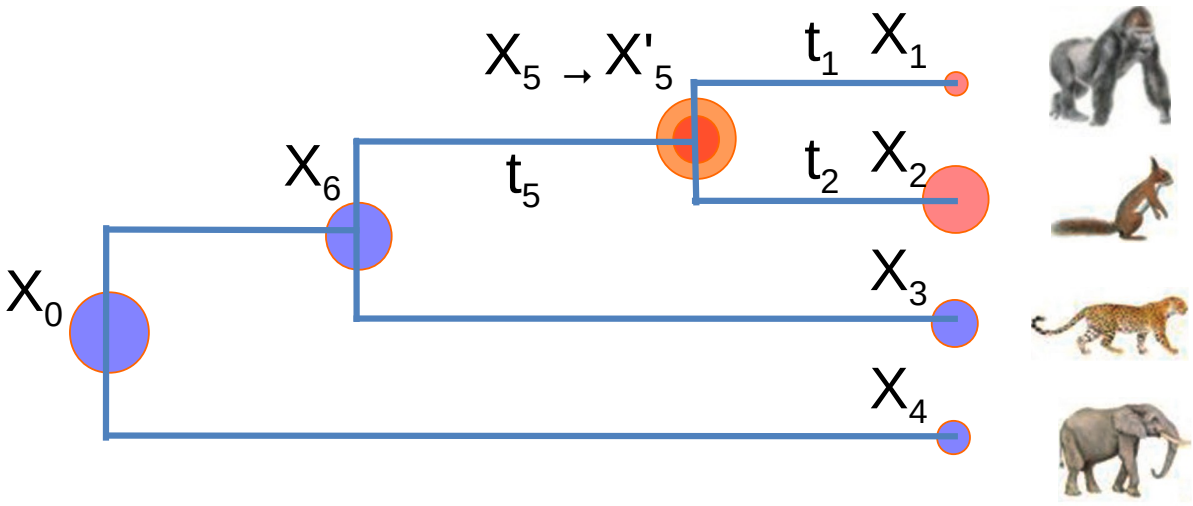
Uninformative priors

- location-invariant, uniform, prior over X_0: $p(X_0) = 1$.
- scale-invariant, log-uniform, prior over σ: $p(\sigma) = \frac{1}{\sigma}$
- frequentist matching (credible intervals = confidence intervals)
- originally due to Harold Jeffreys (1948)

Posterior distribution

$$p(\sigma, X_0, X_{n+1..2n-2} \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..2n-2} \mid X_0, \Psi, \sigma)\, p(X_0)\, p(\sigma)$$

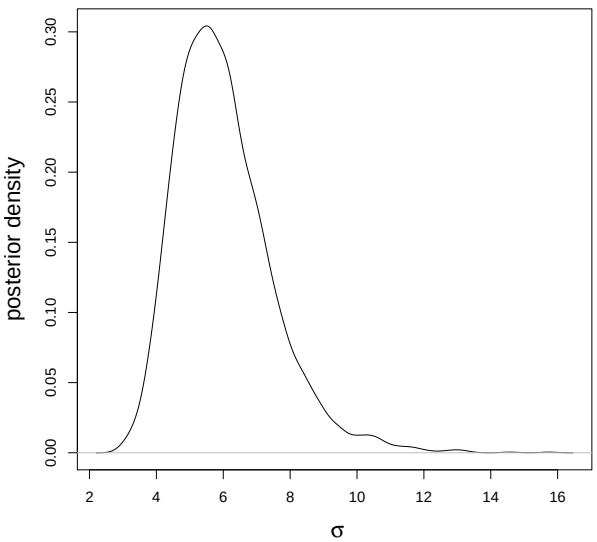
Conditioning on data and MCMC



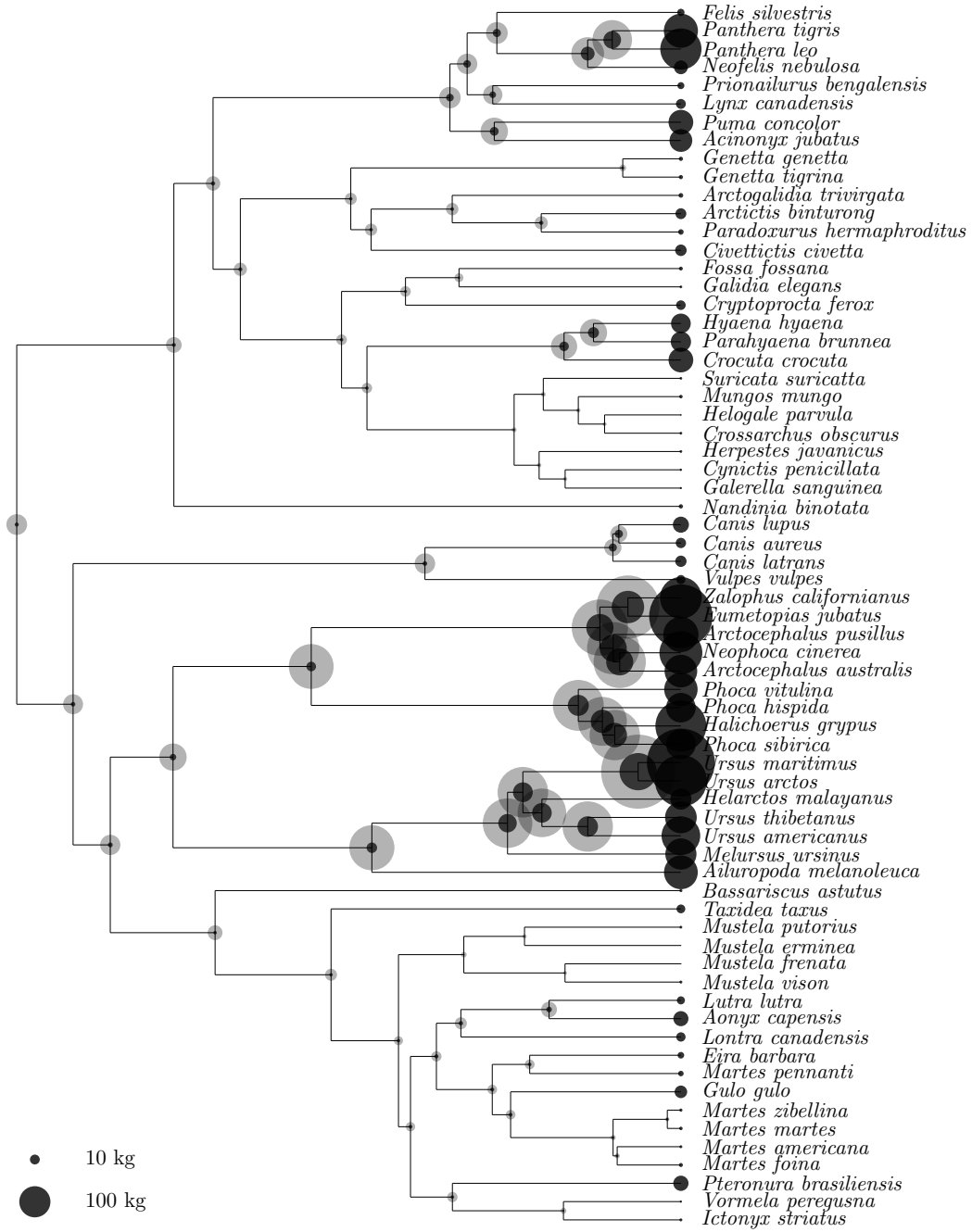
Posterior distribution

$$p(\sigma, X_0, X_{n+1..2n-2} \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..2n-2} \mid X_0, \Psi, \sigma)\, p(X_0)\, p(\sigma)$$

- MCMC over all unclamped variables (ancestral X_j's and σ)
- Metropolis Hastings updates on σ (scaling moves)
- Metropolis Hastings on the X_j's (sliding moves)

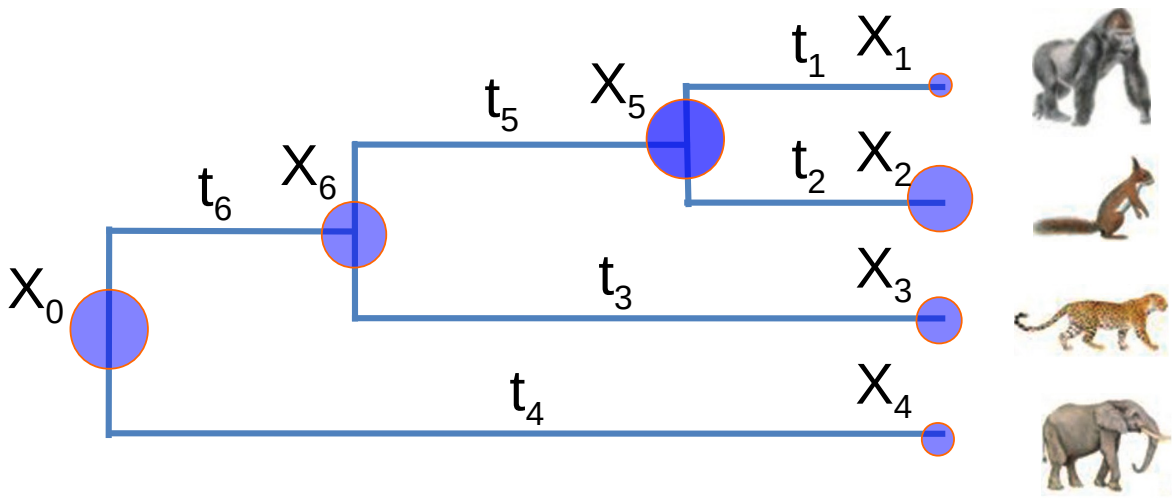


marginal posterior on σ



marginal posterior reconstruction on body mass

Explicit integration over states at interior nodes

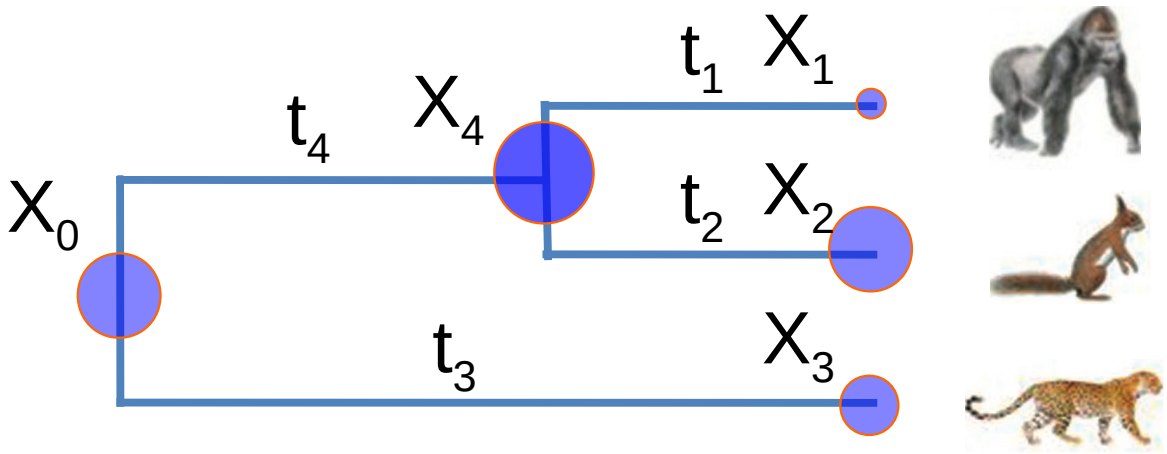


Integrating over internal states ($j = n + 1 \ldots 2n - 2$)

$$p(X_{1..n} \mid X_0, \Psi, \sigma) = \int \ldots \int p(X_{1..2n-2} \mid X_0, T, \sigma) dX_{n+1} \ldots dX_{2n-2}$$

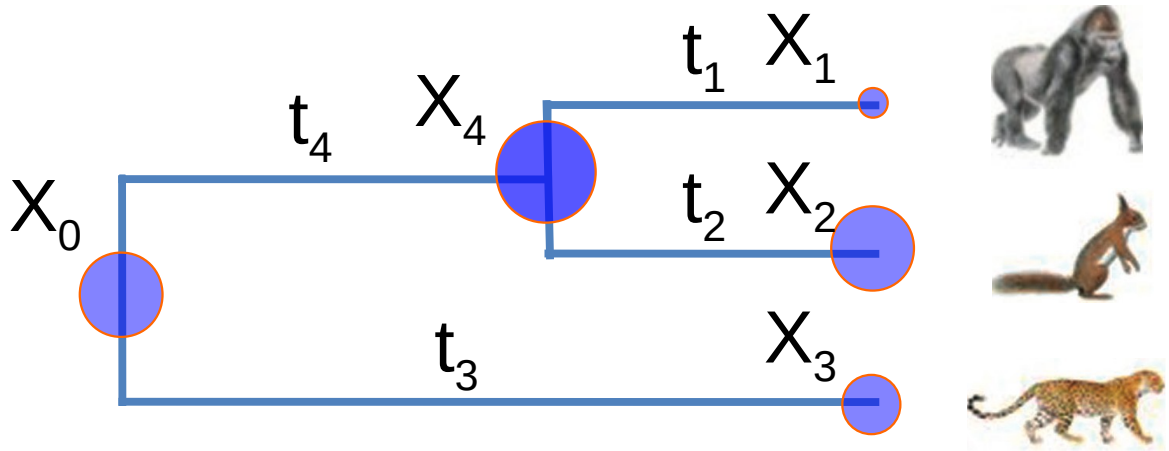
- can be done analytically for Gaussian models
- very simple covariance structure, induced by phylogeny

Covariance between values at the tips



$$
\begin{aligned}
E[X_1 \mid X_0] &= X_0 \\
Var[X_1 \mid X_0] &= \sigma^2(t_1 + t_4) \\
Cov[X_1, X_2 \mid X_0] &= \sigma^2 t_4 \\
Cov[X_1, X_3 \mid X_0] &= 0
\end{aligned}
$$

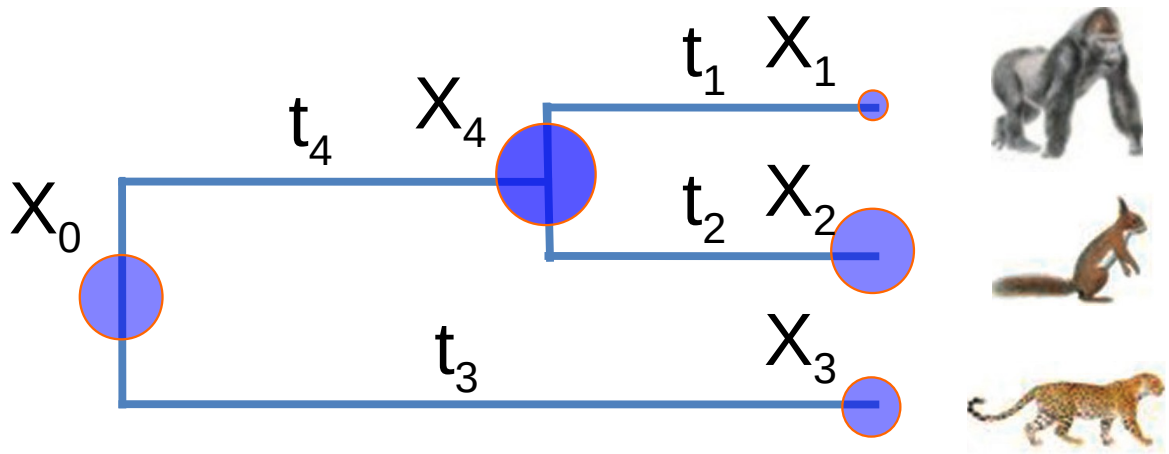
Covariance between values at the tips



Phylogenetic covariance matrix: $C_{ij} = Cov[X_i, X_j]$

$$C = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & \sigma^2(t_1 + t_4) & \sigma^2 t_4 & 0 \\ X_2 & \sigma^2 t_4 & \sigma^2(t_2 + t_4) & 0 \\ X_3 & 0 & 0 & \sigma^2 t_3 \end{array}$$

Integrating the likelihood over interior nodes



$$X_{1..n} \mid X_0, \Psi, \sigma \sim \text{mvNormal}(X_0, C)$$

$$p(X_{1..n} \mid X_0, \Psi, \sigma) \sim \frac{1}{\sqrt{(2\pi)^n |C|}} e^{-\frac{1}{2}(X_{1..n}-X_0)' C^{-1} (X_{1..n}-X_0)}$$

integrating over uniform prior for X_0 ($p(X_0) = 1$)

$$p(X_{1..n} \mid \Psi, \sigma) = \int_{-\infty}^{+\infty} p(X_{1:n} \mid X_0, \Psi, \sigma^2) p(X_0) dX_0$$

Integrating versus sampling

Augmented model

$$p(\sigma, X_0, X_{n+1..2n-2} \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..2n-2} \mid X_0, \Psi, \sigma)\, p(X_0)\, p(\sigma)$$

Integrated model

$$p(\sigma \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..n} \mid \Psi, \sigma)\, p(\sigma)$$

Equivalence principle

- the two settings are mathematically equivalent
- computational efficiency: depends on the context
- data-augmentation very useful when integrals are not analytical

Other models of trait evolution

Gaussian models

- Brownian motion with systematic trend (drift)
- Ornstein-Uhlenbeck process

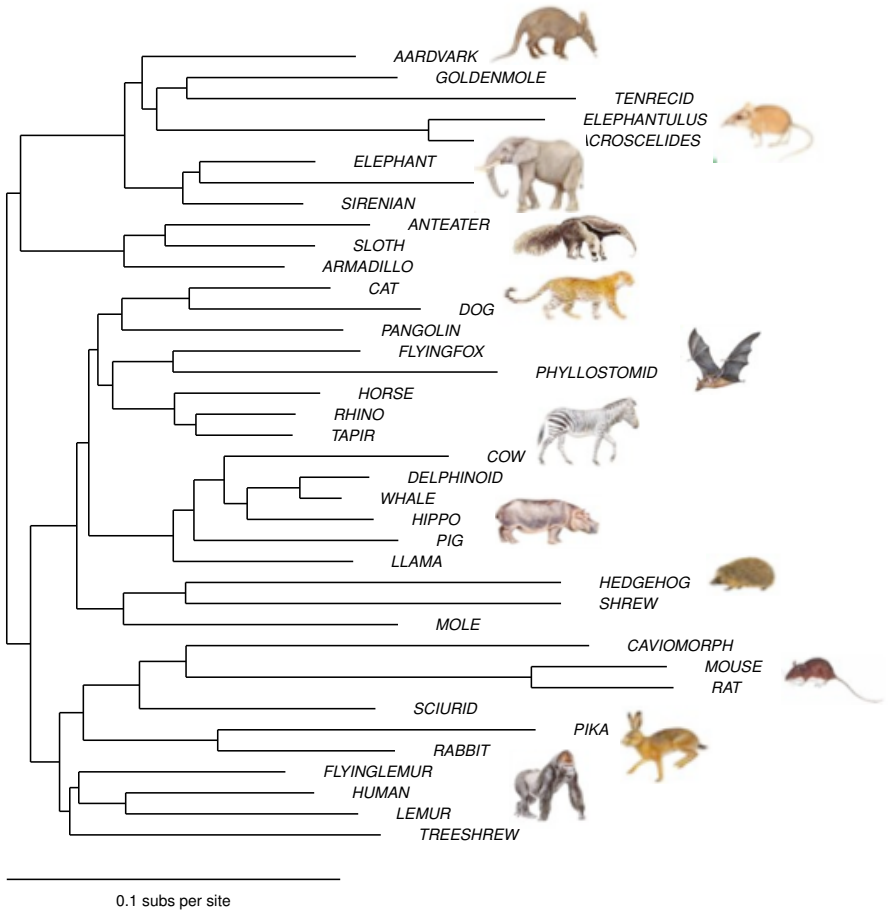
Non-Gaussian models

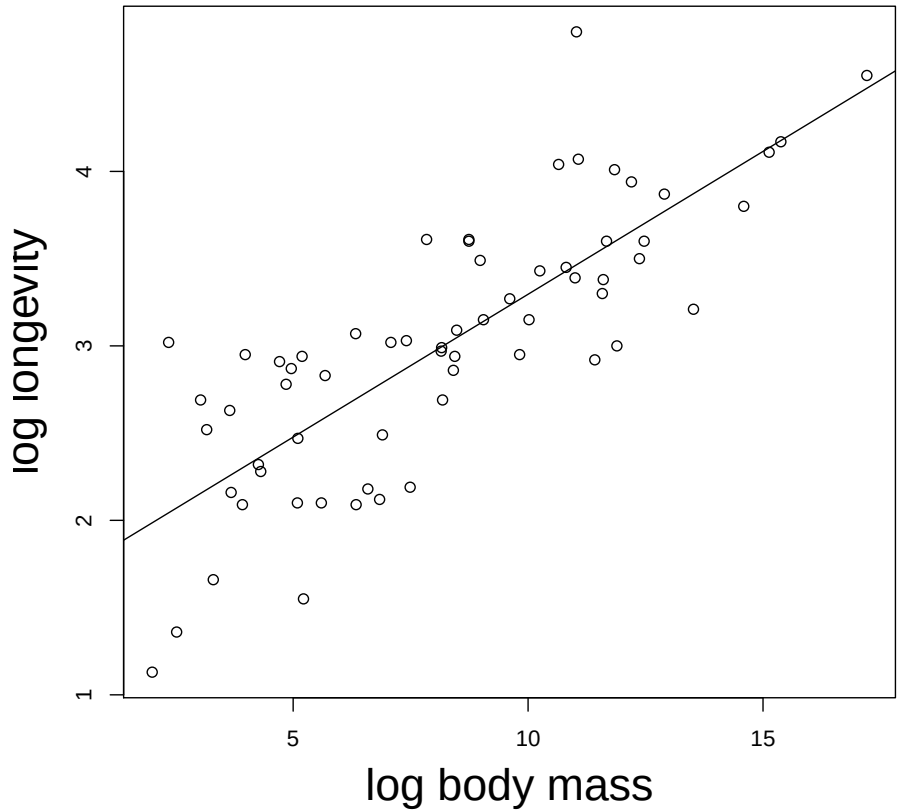
- Levy processes (jumps, increments of infinite variance)

Compound processes

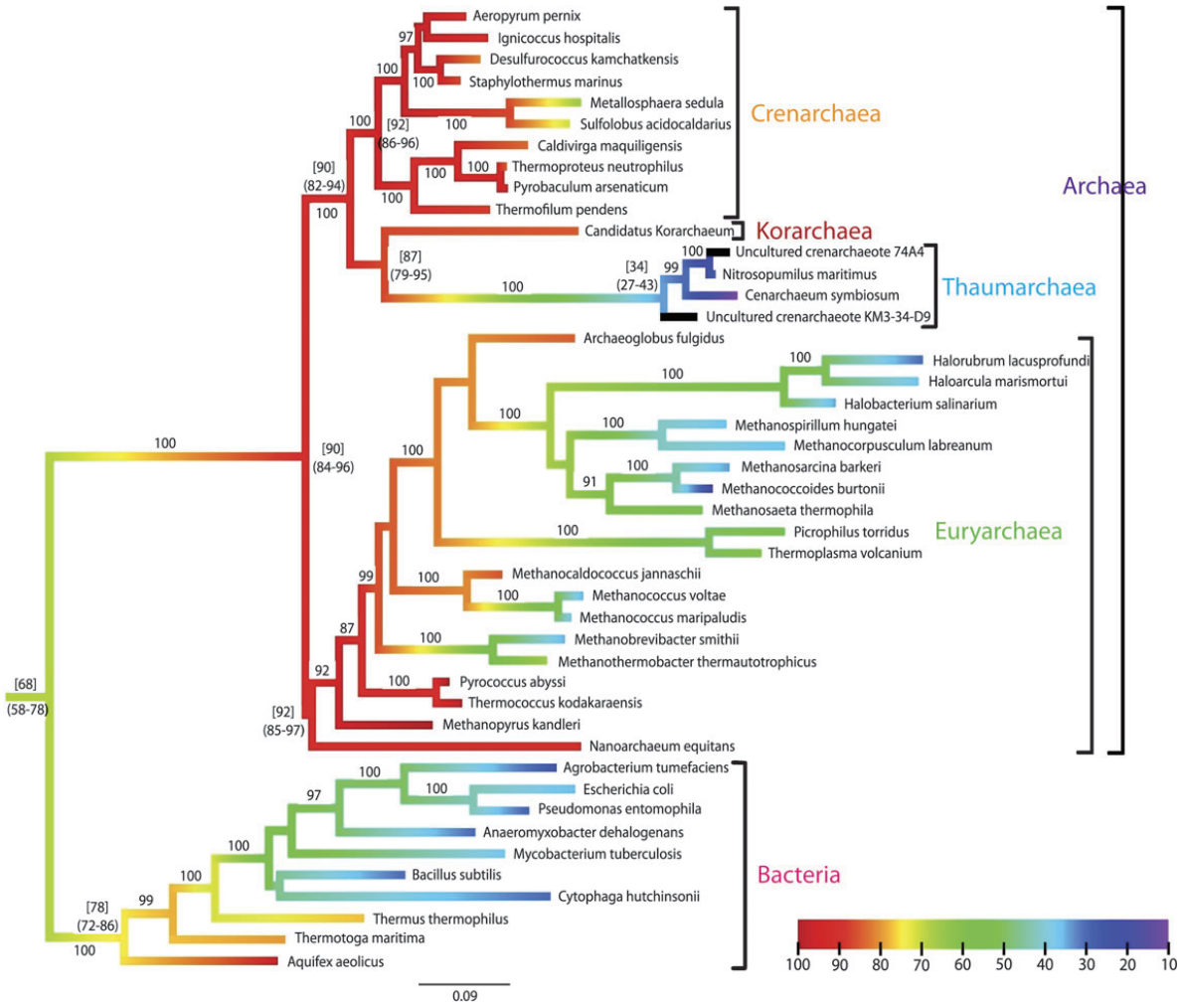
- Brownian whose rate changes at discrete points
- OU with optimum and/or rates changing at discrete points
- change points themselves from Poisson process

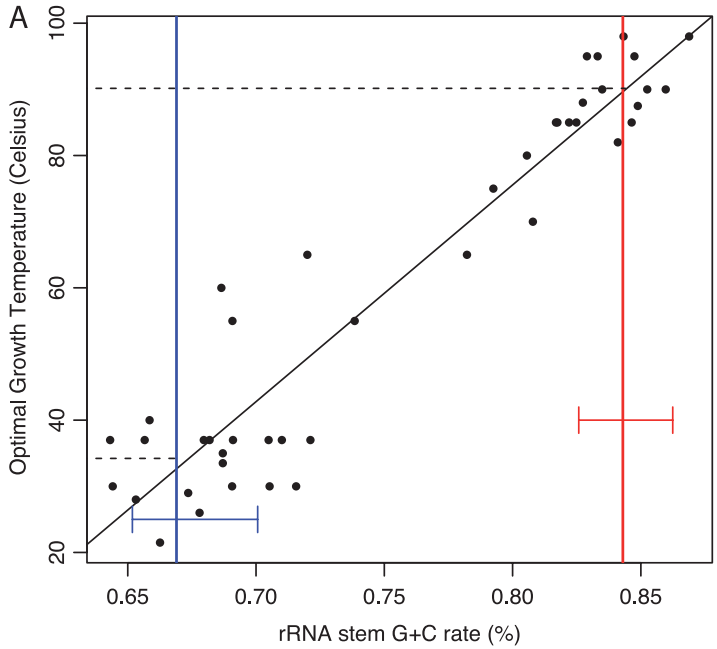
Multiple traits – correlated evolution





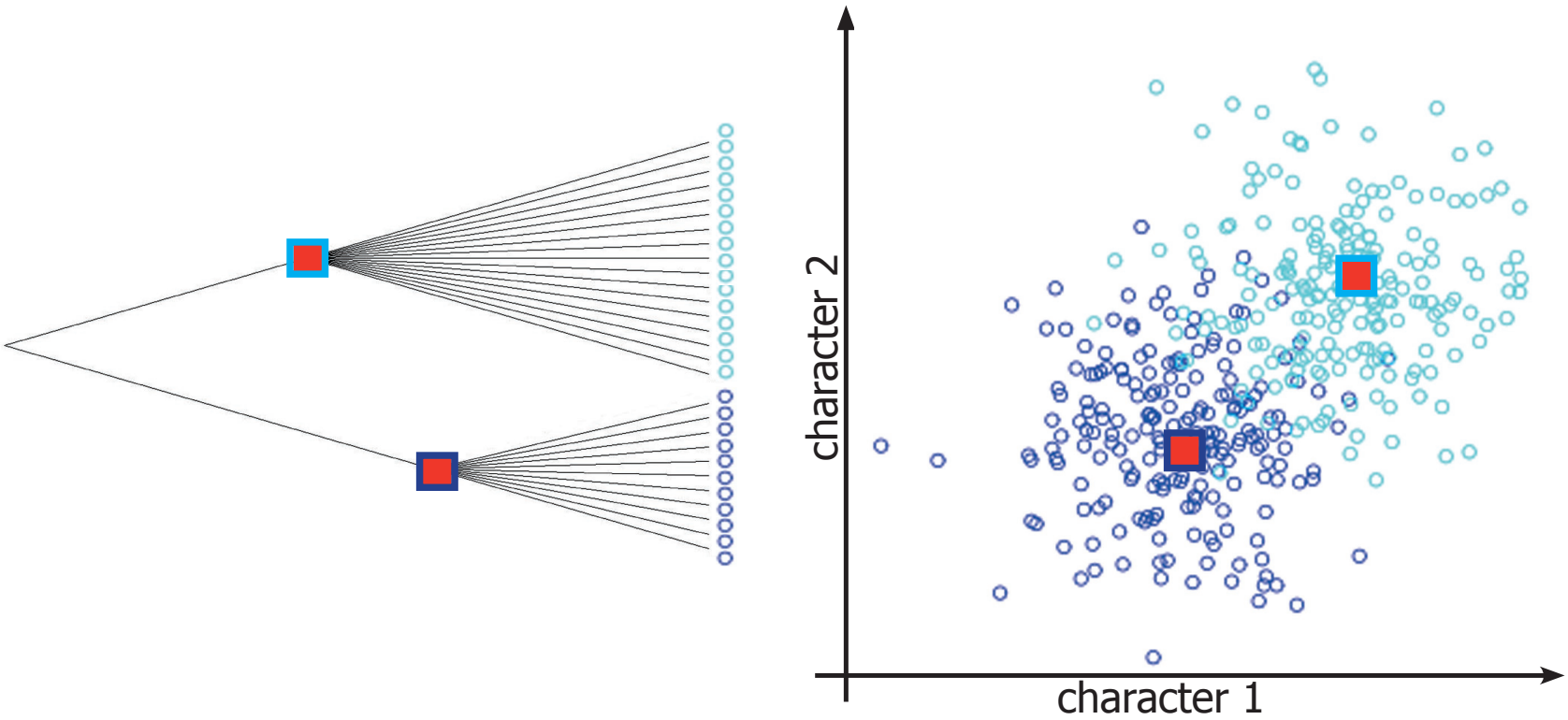
Ancestral growth temperatures inferred using rRNA





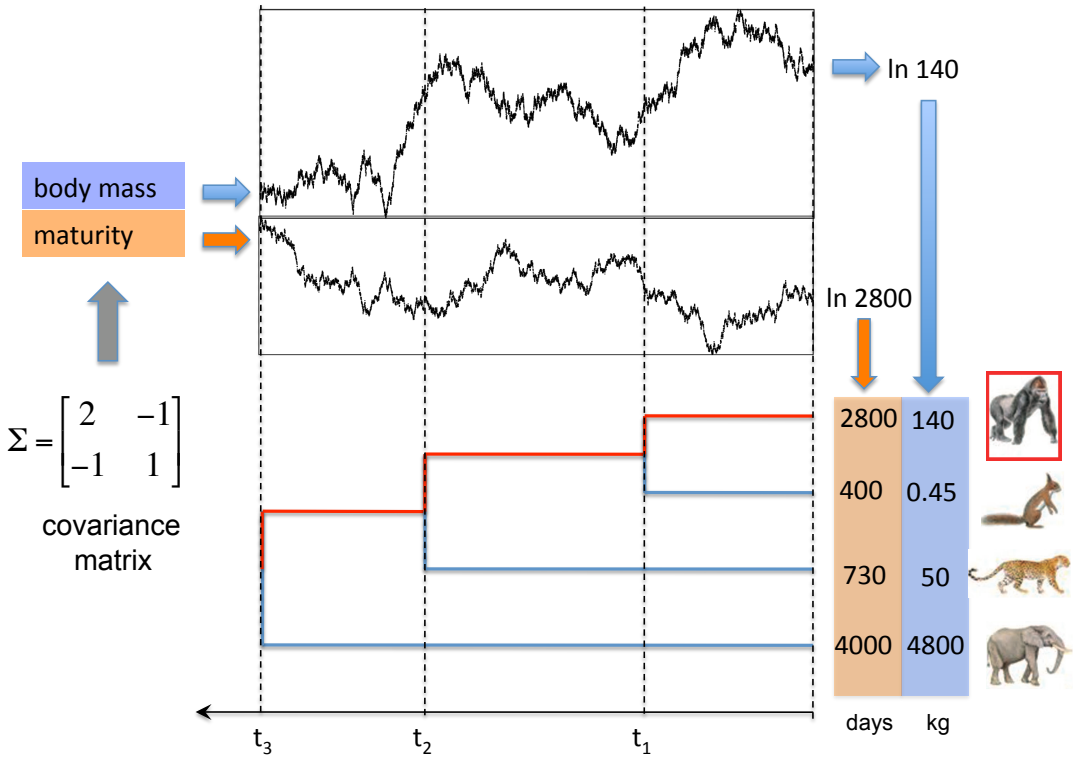
Groussin et al, 2011, Mol Biol Evol 28:2661

The problem of phylogenetic inertia



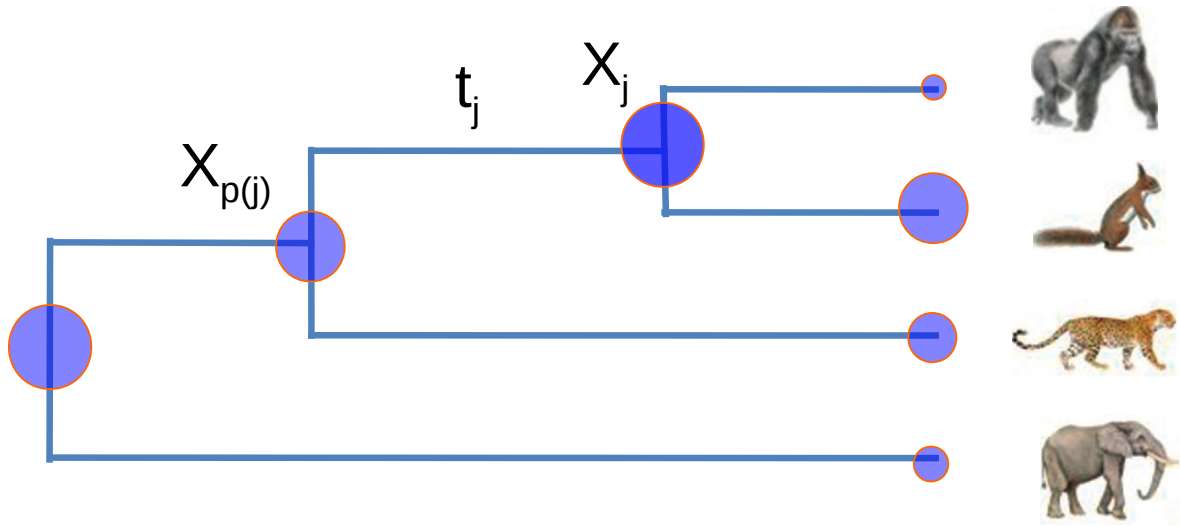
Felsenstein, 1985, Am Nat 125:1

Joint evolutionary process



- Assume 2 traits follow bivariate Brownian motion
- data $X = (x_{jk})$, $j = 1..n$, $k = 1, 2$
- unknown covariance matrix Σ (3 independent parameters)
- estimate Σ_{12}, test whether $\Sigma_{12} < 0$, etc

Multivariate Brownian model of trait evolution

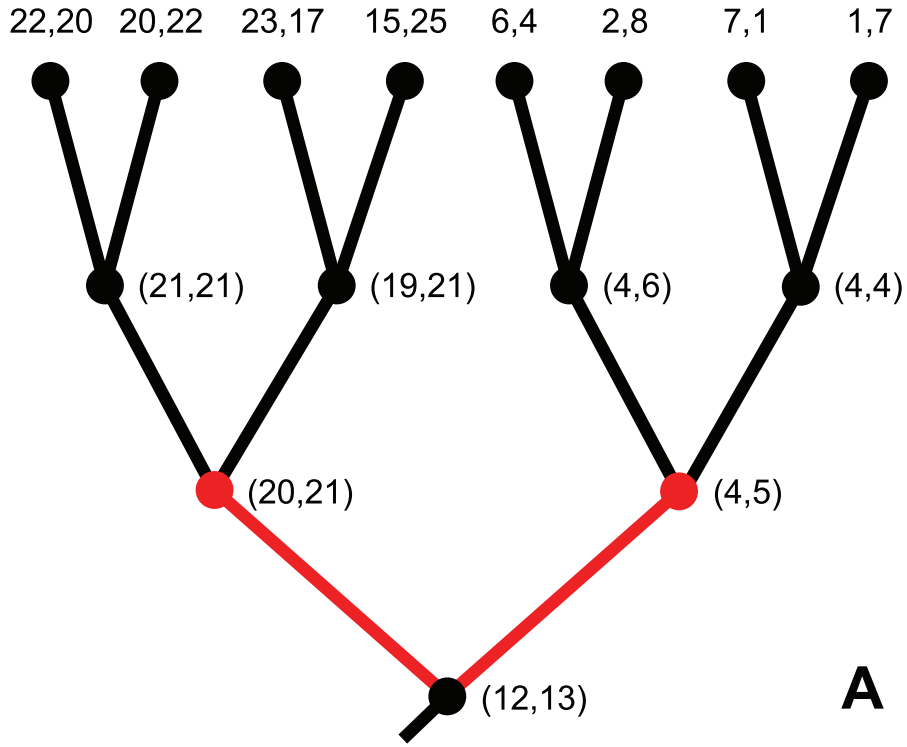


Markov decomposition

$$p(X_{1..2n-2} \mid X_0, \Psi, \Sigma) = \prod_{j=1}^{2n-2} p\left(X_j \mid X_{p(j)}, t_j, \Sigma\right)$$

$$X_j \mid X_{p(j)}, t_j, \Sigma \sim \text{mvNormal}\left(X_{p(j)},\ t_j\Sigma\right)$$

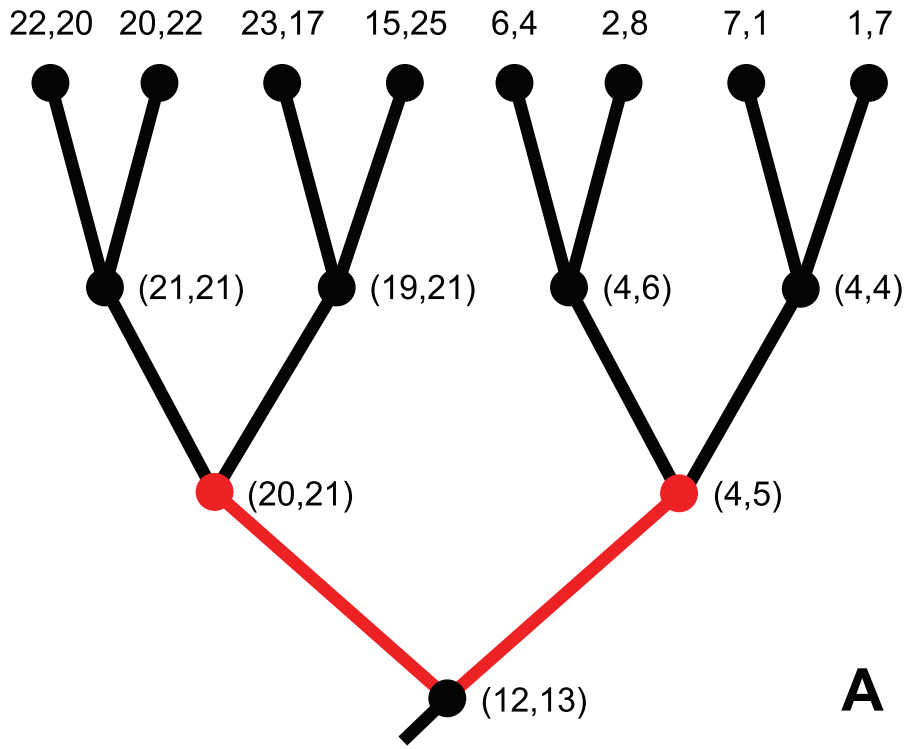
The method of independent contrasts



Algorithm

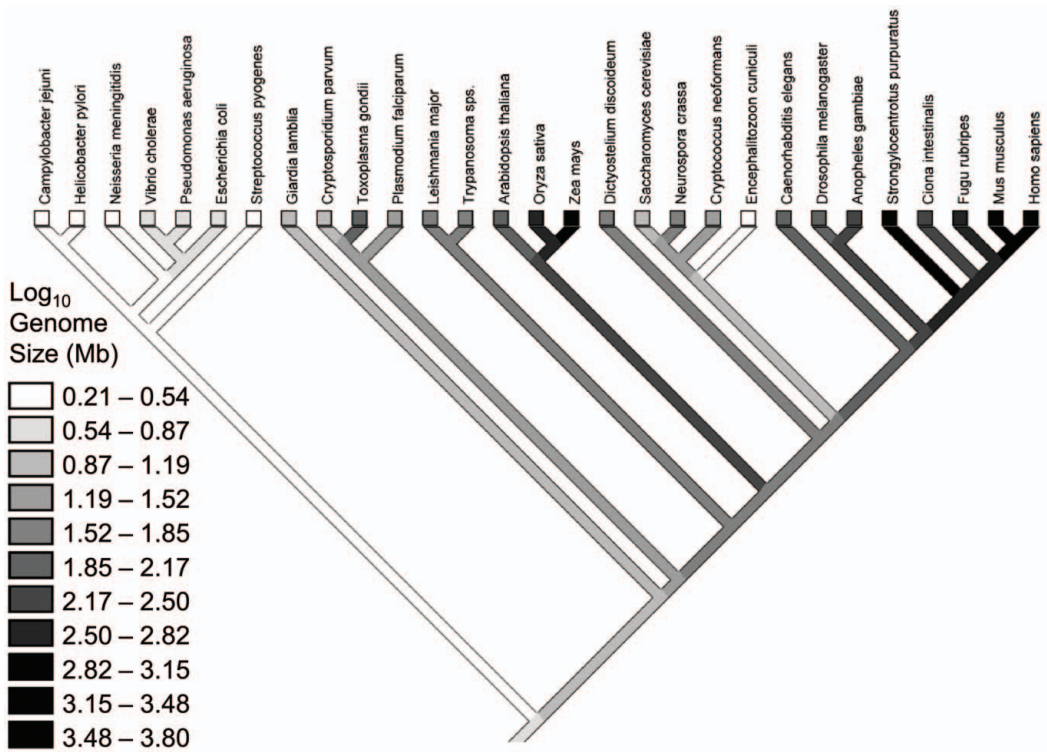
- take two sister species i and j, with traits X_i and X_j
- t: time since their last common ancestor
- $\Delta X = X_j - X_i \sim Normal(0, 2t\Sigma)$.
- define normalized *contrast* $\Delta Y = \Delta X/\sqrt{2t}$: $\Delta Y \sim Normal(0, \Sigma)$.
- do it for all tips and then recursively up to the root

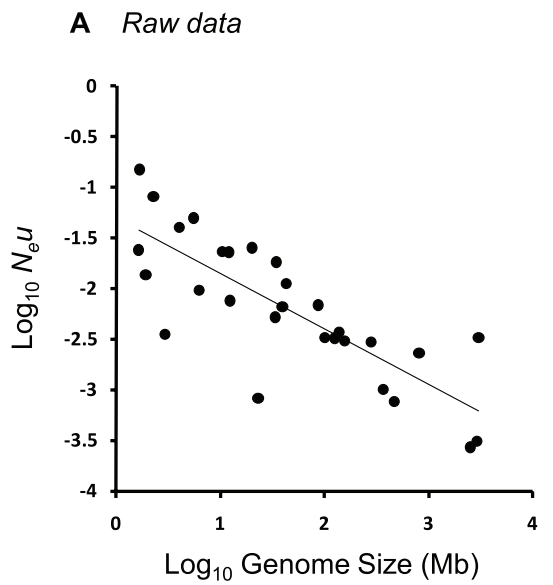
The method of independent contrasts

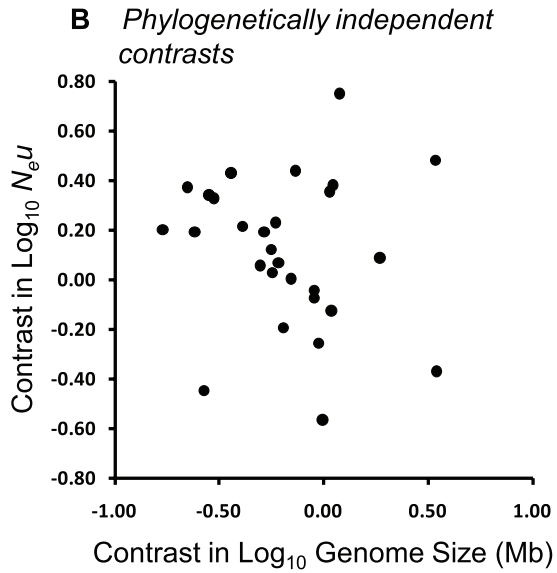


- P taxa $\rightarrow P-1$ normalized contrasts ΔY_j, for $j = 1..P-1$
- normalized contrasts are iid: $\Delta Y_j \sim Normal(0, \Sigma)$
- contrasts are statistically independent
- equivalent to asking whether traits show correlated *variations*

Genome size and effective population size revisited

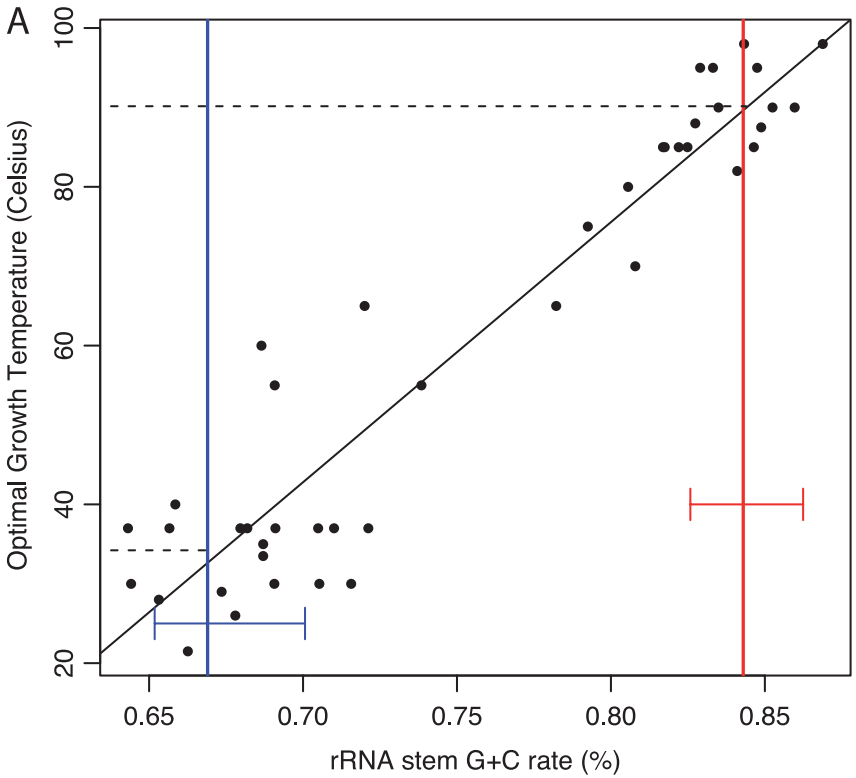




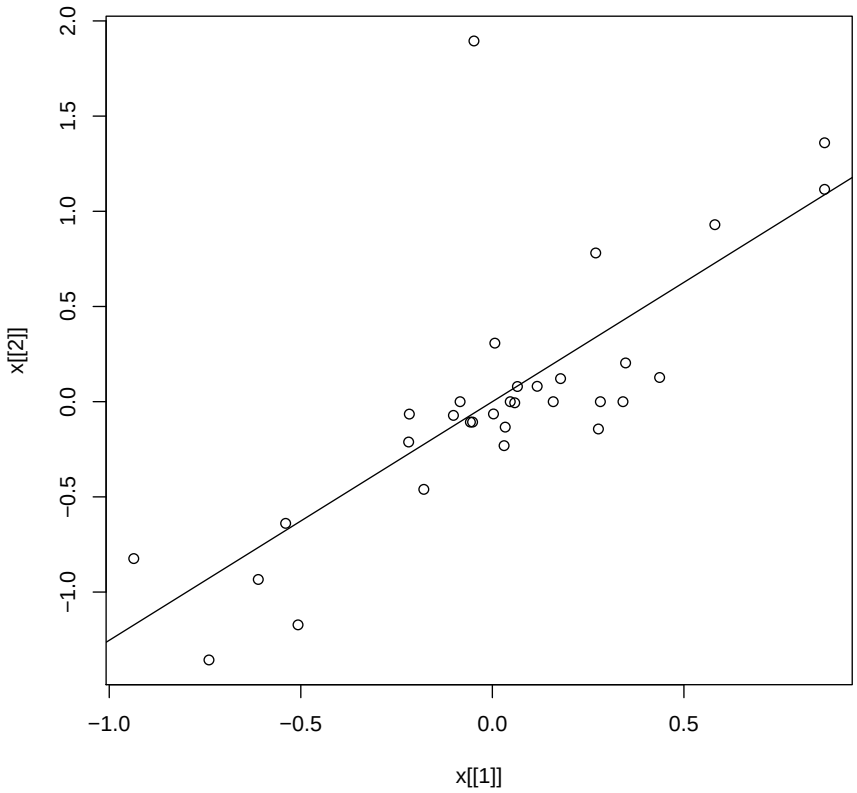


Whitney and Garland Jr, 2010, PLoS Genetics 6:e1001080

Phylogenetically-corrected regression in Archaea

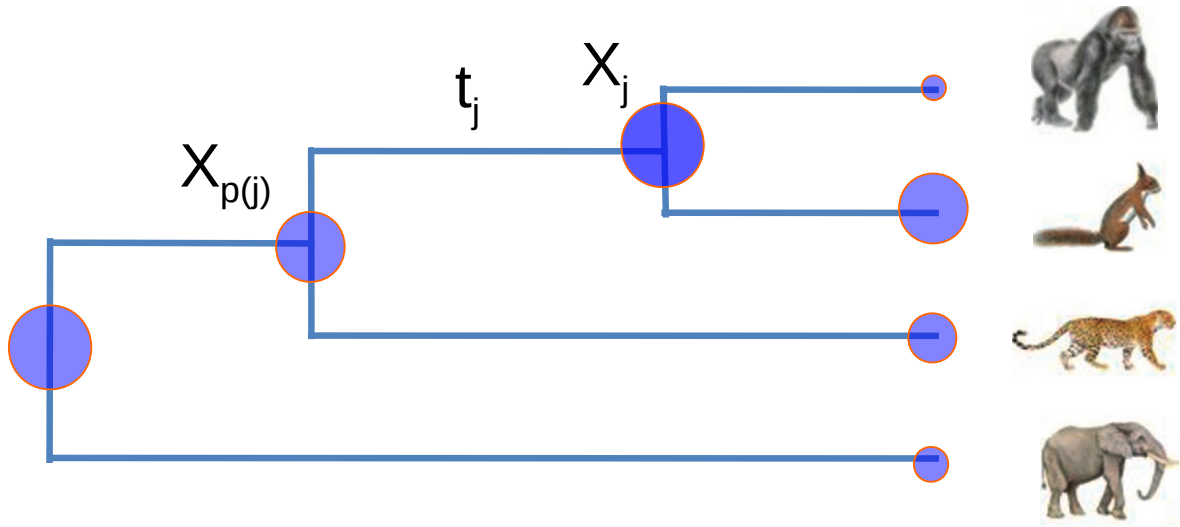


raw regression



normalized contrasts

Multivariate Brownian model of trait evolution



Likelihood

$$p(X_{1..2n-2} \mid X_0, \Psi, \Sigma) \quad = \quad \prod_{j=1}^{2n-2} p\left(X_j \mid X_{p(j)}, t_j, \Sigma\right)$$

posterior

$$p(\Sigma, X_0, X_{n+1..2n-2} \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..2n-2} \mid X_0, \Psi, \Sigma)\, p(X_0)\, p(\Sigma)$$

Prior distributions over the covariance matrix

- V: $K \times K$ covariance matrix
- draw $Y \sim \text{mvNormal}(0, V)$
- by definition, $Cov[Y_i, Y_j] = V_{ij}$

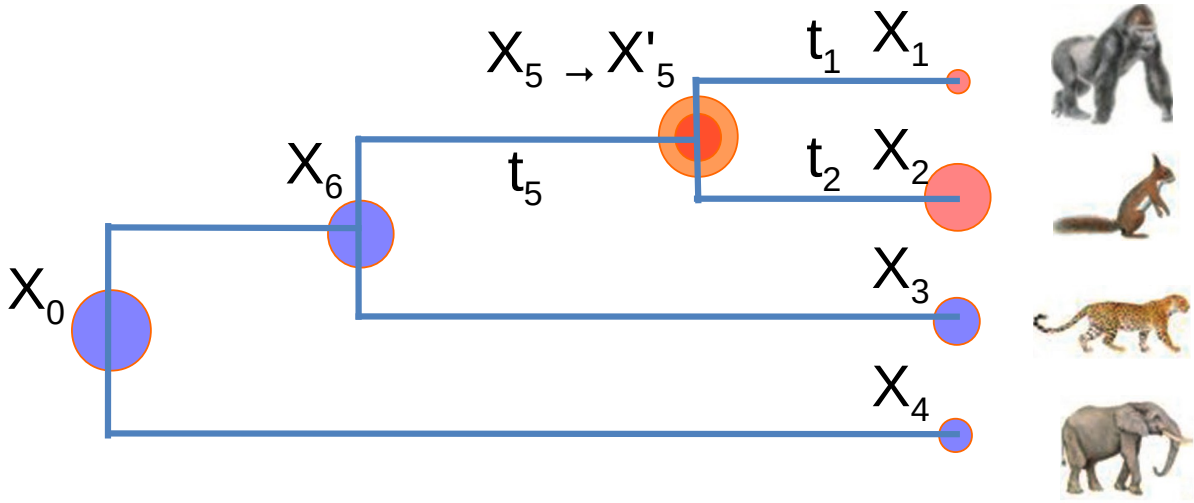
Wishart distribution

- draw p independent samples from $\sim \text{mvNormal}(0, V)$
- calculate their *empirical* covariance matrix $\hat{V}$
- define $\Omega = p\hat{V}$
- then $\Omega \sim \text{Wishart}(V, p)$

Inverse Wishart distribution

- draw $\Omega \sim \text{Wishart}(V, p)$
- set $\Sigma = \Omega^{-1}$ and $W = V^{-1}$
- then $\Sigma \sim \text{invWishart}(W, p)$

Conditioning on data and MCMC

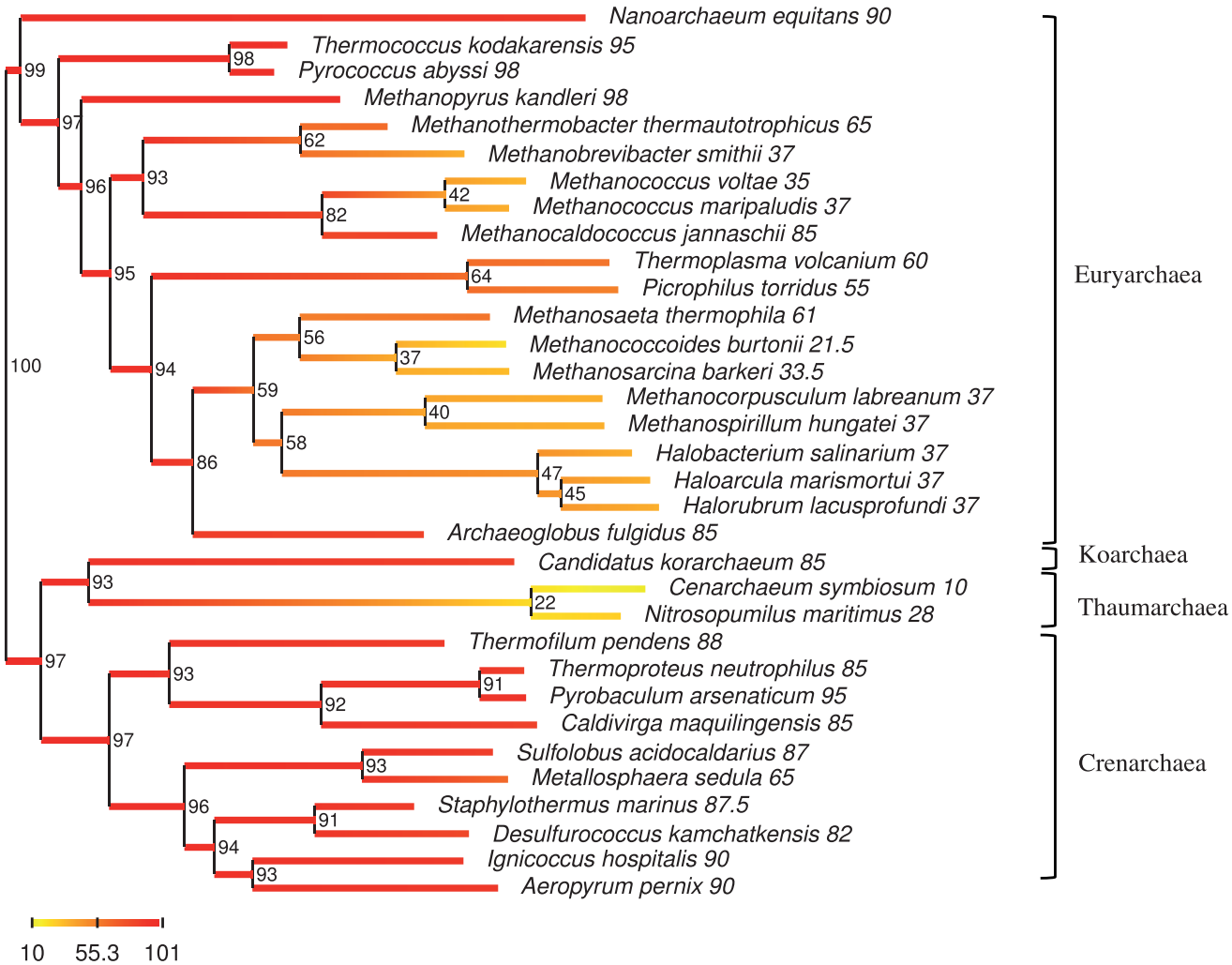


Posterior distribution

$$p(\Sigma, X_0, X_{n+1..2n-2} \mid X_{1..n}, \Psi) \quad \propto \quad p(X_{1..2n-2} \mid X_0, \Psi, \Sigma)\, p(X_0)\, p(\Sigma)$$

- MCMC over all unclamped variables
- clamped variables include traits in extant taxa
- but also potentially traits in ancestors

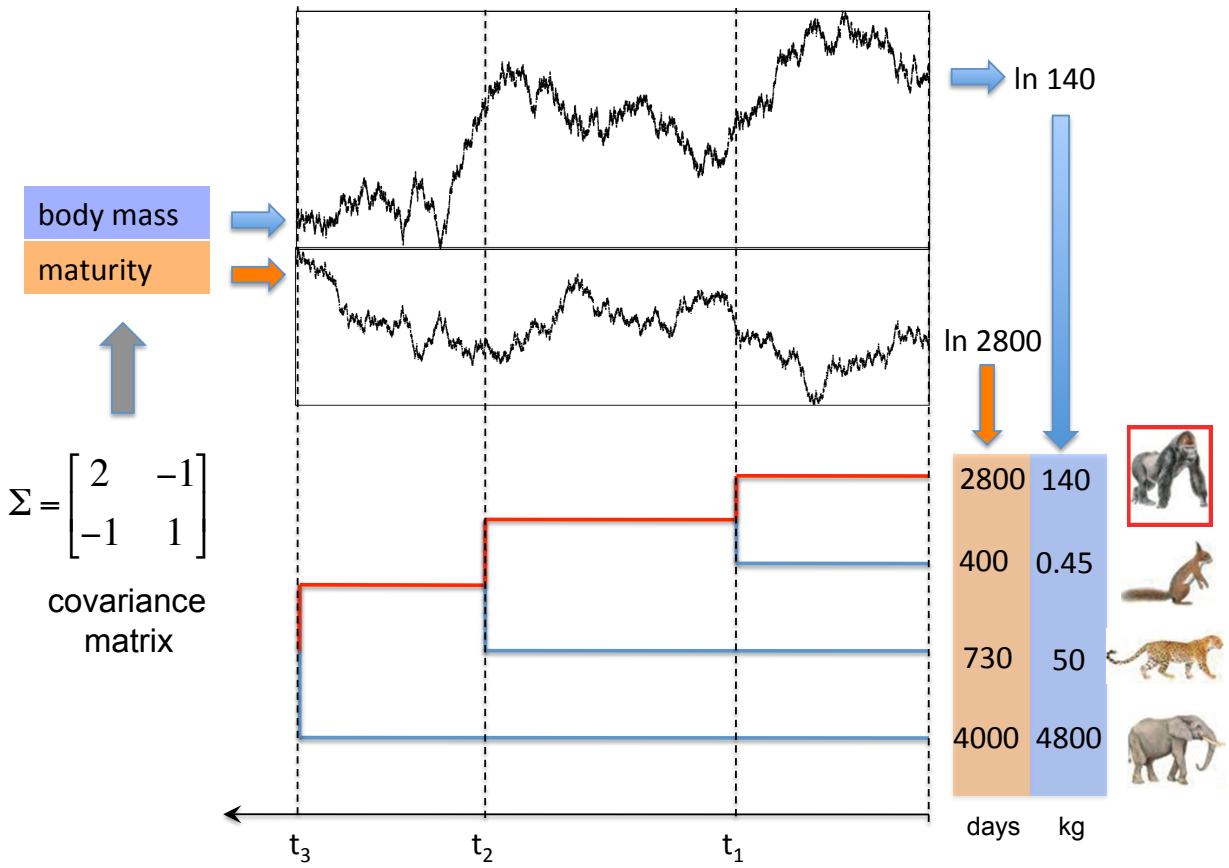
archaeal rRNA dataset



Inferred temperature for archaeal ancestor

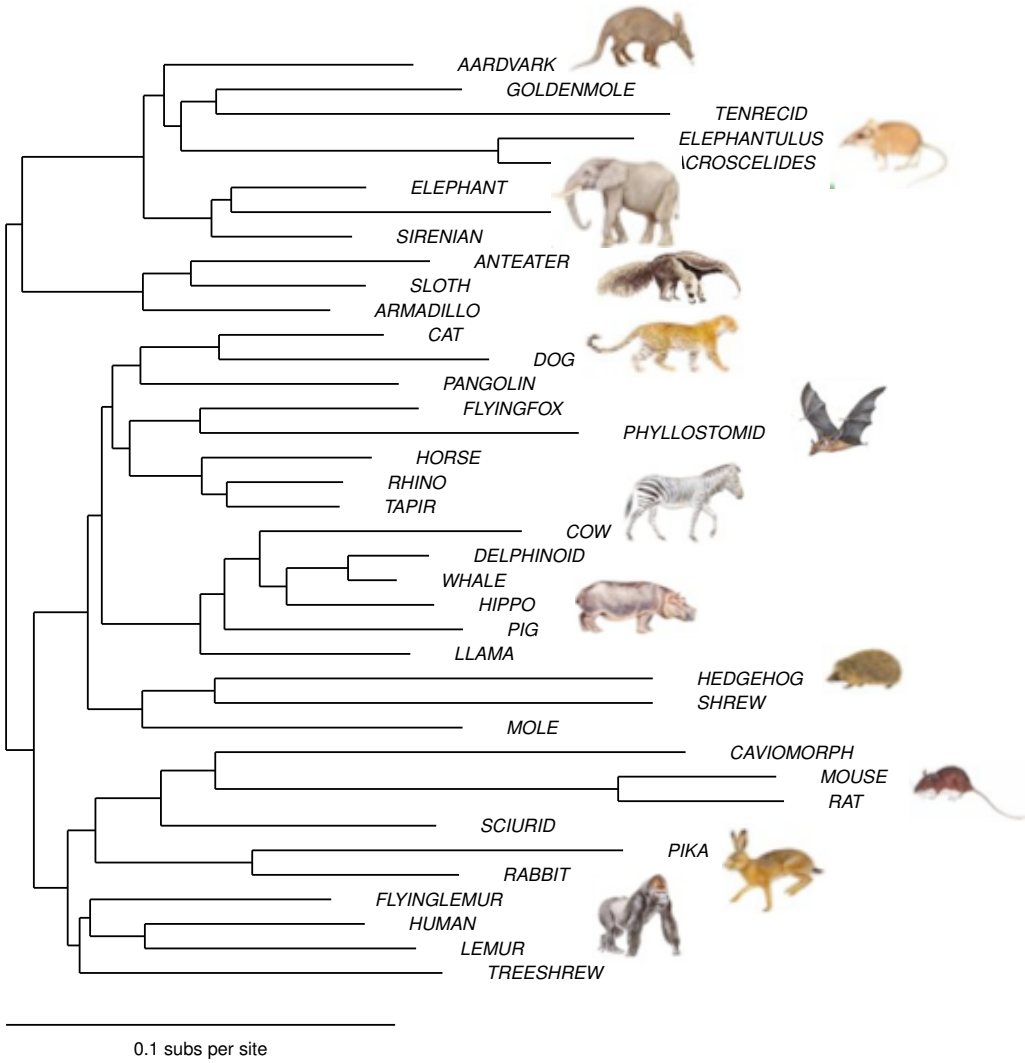
- 95% credible interval: (91,110) Celsius.
- without molecular information: (60,96) Celsius

The comparative method – Summary



- independent contrast $\Longleftrightarrow$ traits follow bivariate Brownian motion
- more generally: statistical models of the evolutionary *processes*
- a large variety of questions: bursts, trends, jumps, correlations.

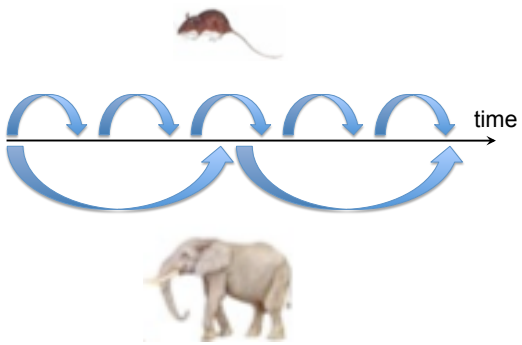
Variation of the substitution rate among lineages



Variation of the substitution rate among lineages

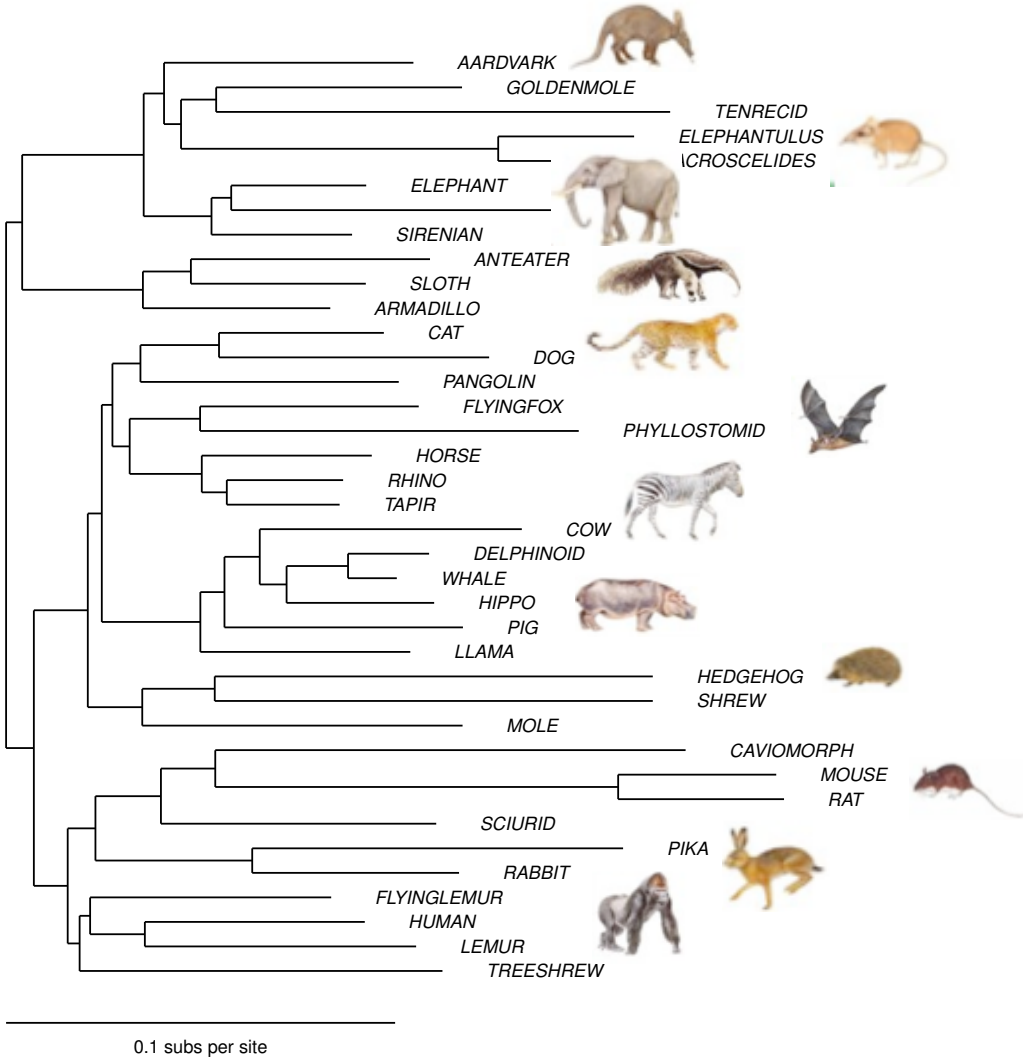
Possible causes

- generation-time effect

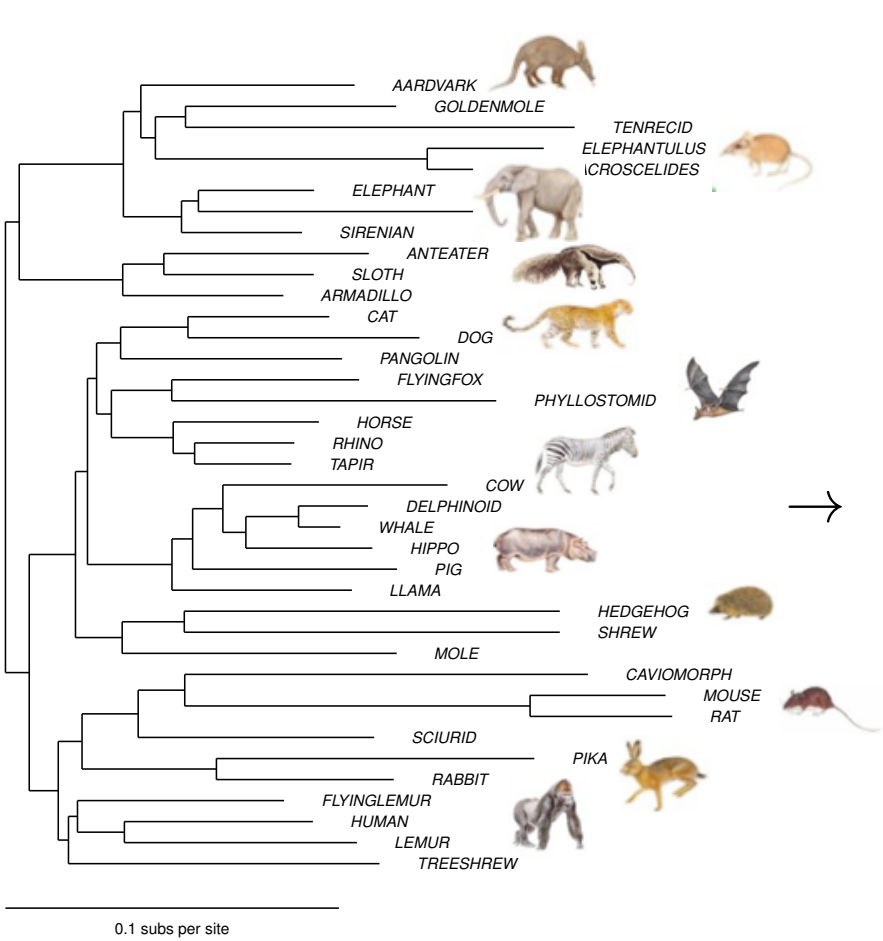


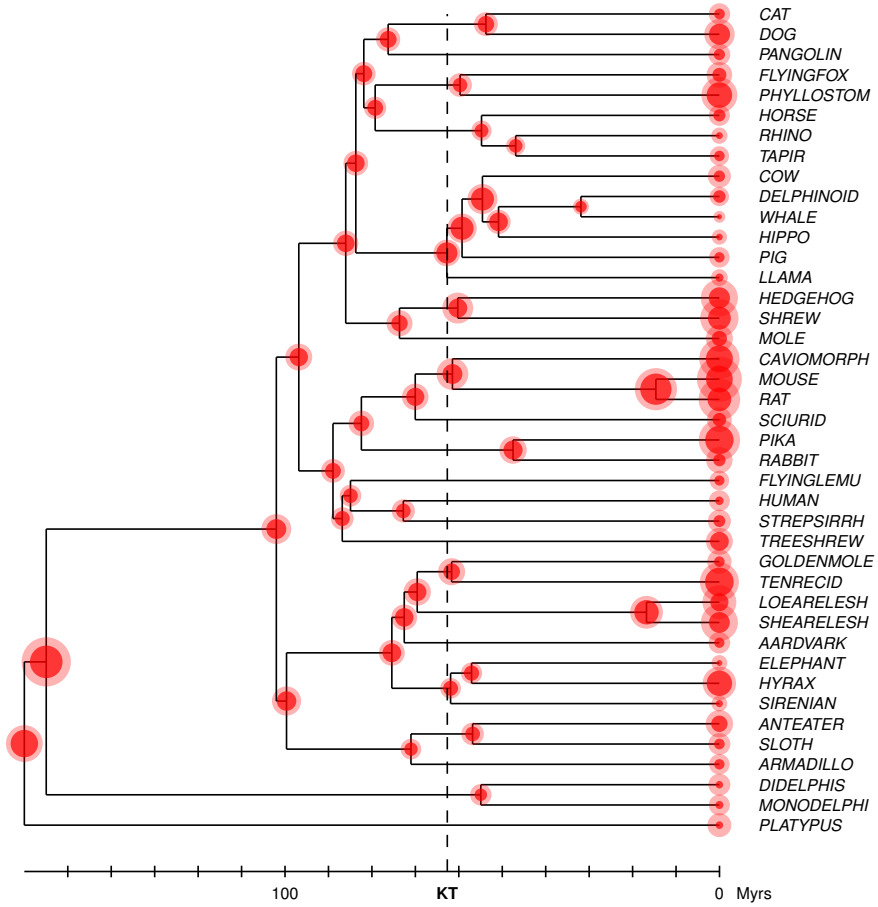
- metabolic rate effects
- selection for longevity

(reviewed in Lanfear et al, 2010)



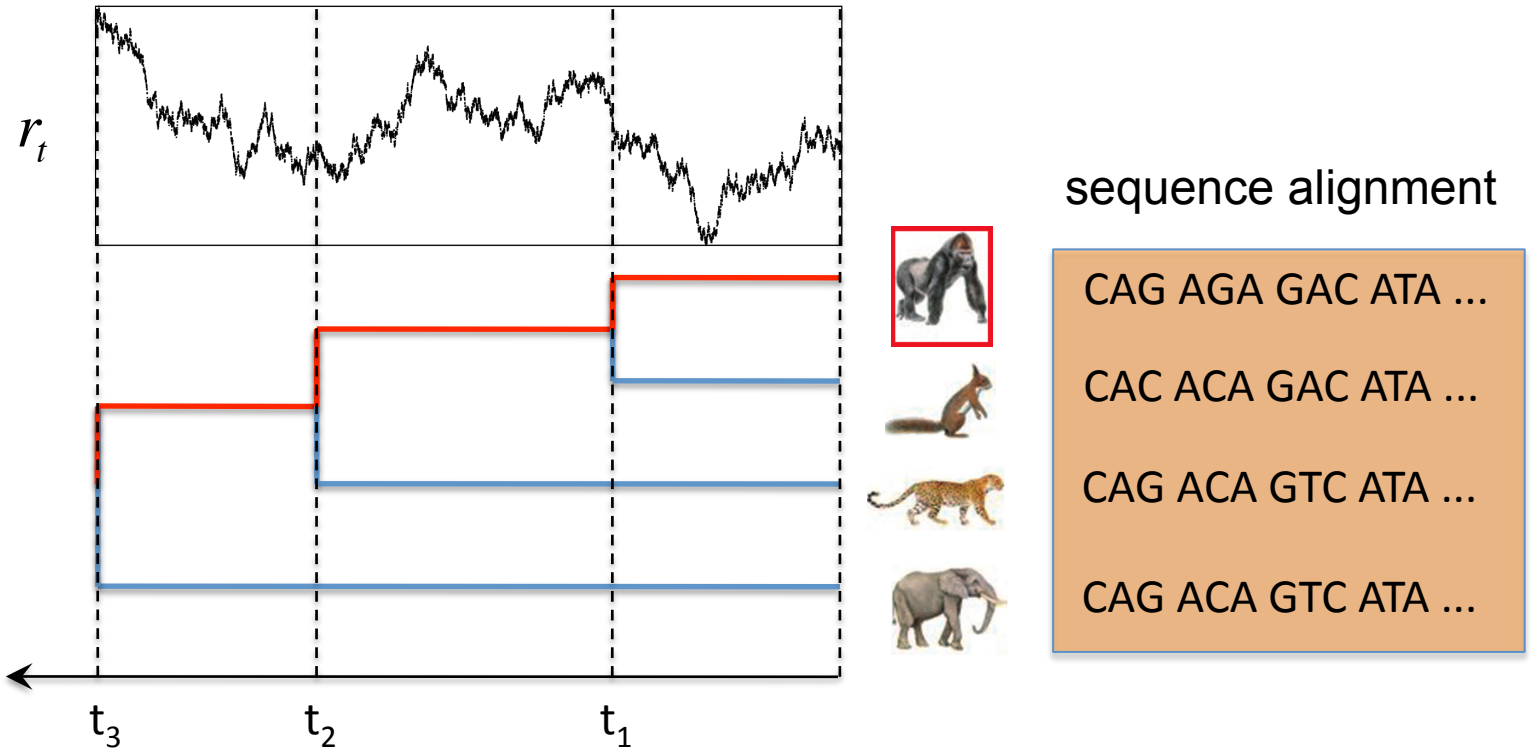
The relaxed molecular clock





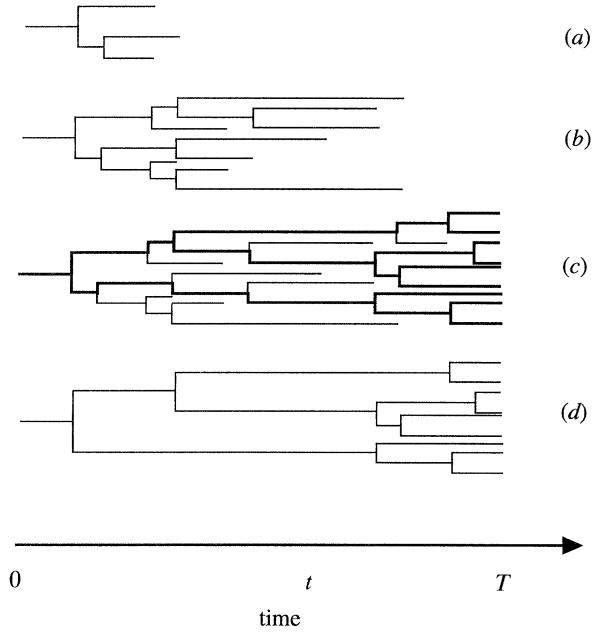
Branch lengths = times x rates

The Brownian relaxed molecular clock



- substitutions occur at rate r_t
- In r_t modeled as Brownian motion along branches
- Brownian model induces rate *autocorrelation* across branches
- joint estimation of rates and times by Bayesian MCMC

Diversification process

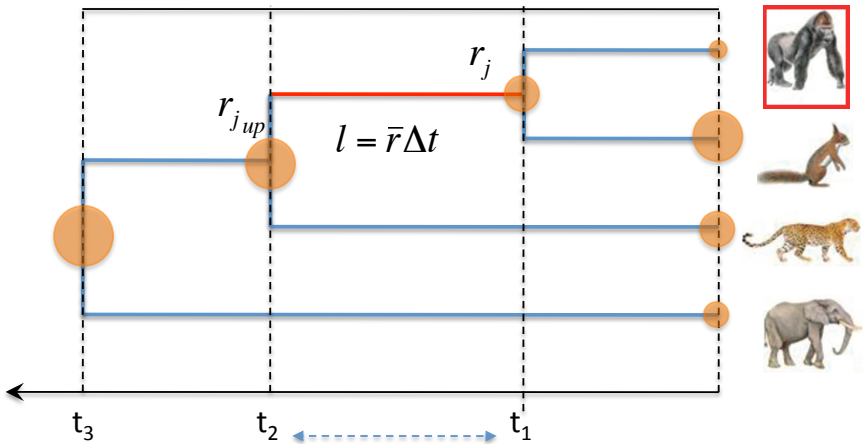


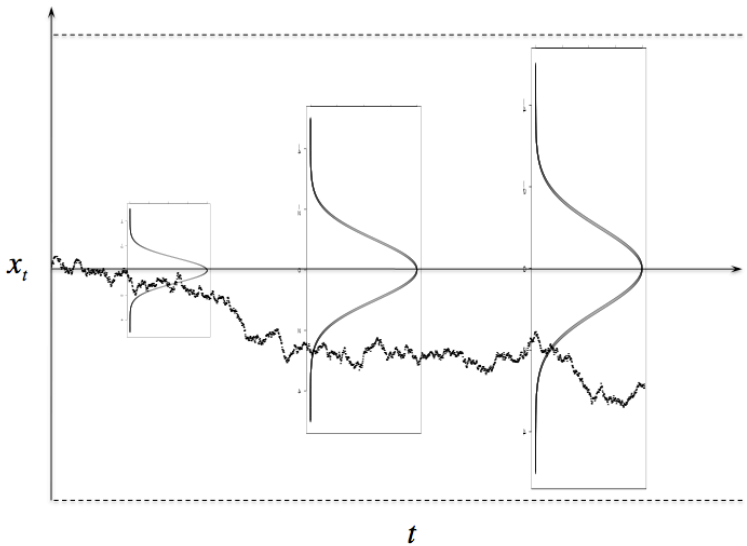
Nee et al, 1994

birth-death with subsampling

- speciation rate λ, extinction rate μ, sampling fraction ρ
- Ψ: time-calibrated phylogeny
- gives you a probability distribution on times: $p(\Psi \mid \lambda, \mu, \rho)$

Brownian process





$$x_t = \ln r_t$$
$$x_t \sim N(x_0, \sigma^2 t)$$

gives you a probability distribution on rates: $p(x \mid x_0, \Psi, \sigma)$

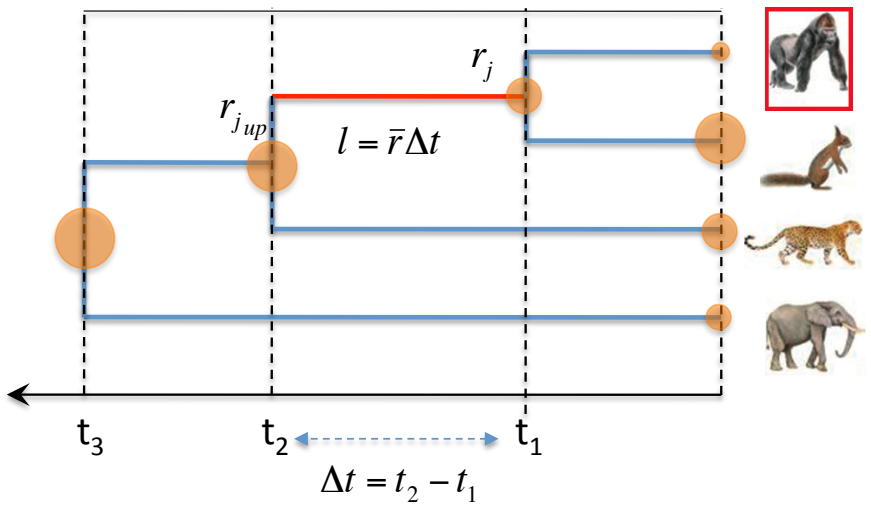
Model of sequence evolution by point substitutions

Substitution rate matrix Q (4 x 4)

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \frac{\gamma}{2} & \kappa\frac{\gamma}{2} & \frac{1-\gamma}{2} \\ C & \frac{1-\gamma}{2} & - & \frac{\gamma}{2} & \kappa\frac{1-\gamma}{2} \\ G & \kappa\frac{1-\gamma}{2} & \frac{\gamma}{2} & - & \frac{1-\gamma}{2} \\ T & \frac{1-\gamma}{2} & \kappa\frac{\gamma}{2} & \frac{\gamma}{2} & - \end{array}$$

- κ: transition-transversion ratio
- γ: equilibrium GC (GC^*)

Substitution process (rate matrix Q)



sequence alignment

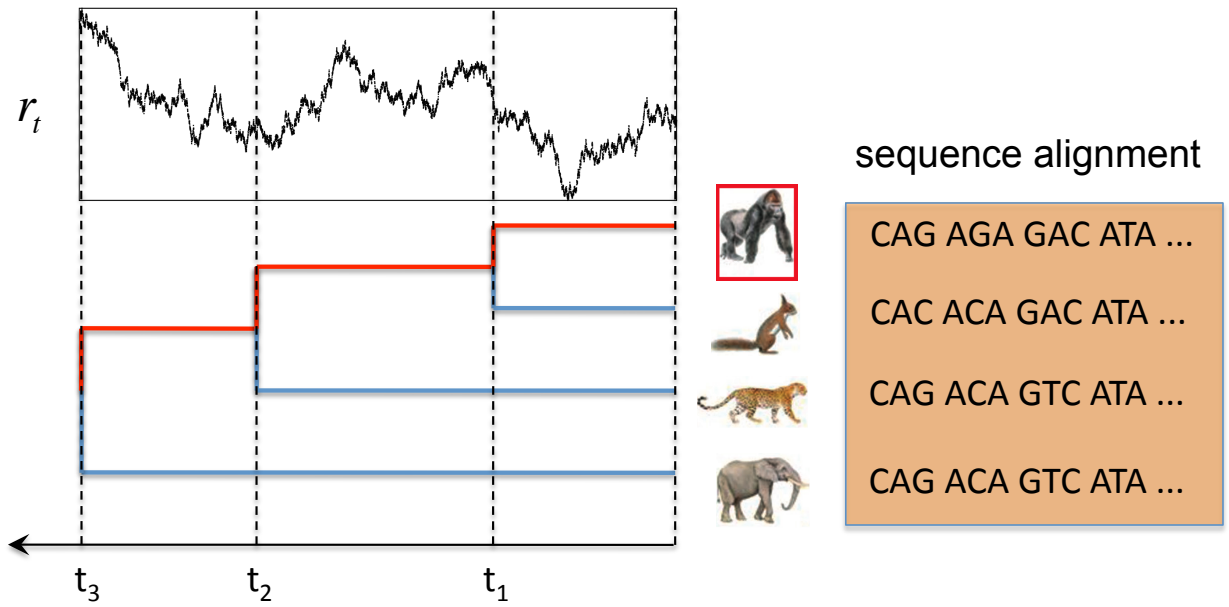
CAG AGA GAC ATA ...
CAC ACA GAC ATA ...
CAG ACA GTC ATA ...
CAG ACA GTC ATA ...

- branch length approximated as:

$$l = \bar{r}\Delta t, \quad \text{where} \quad \bar{r} = \frac{r_j + r_{j_{up}}}{2}$$

- length l: expected number of point substitutions along the branch
- gives you a probability distribution on sequences: $p(D \mid r, \Psi, Q)$

Complete model

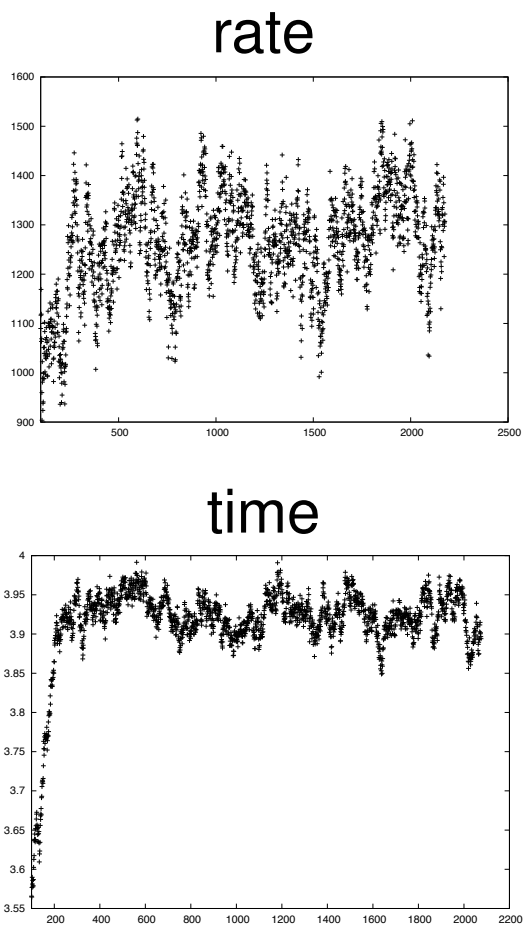


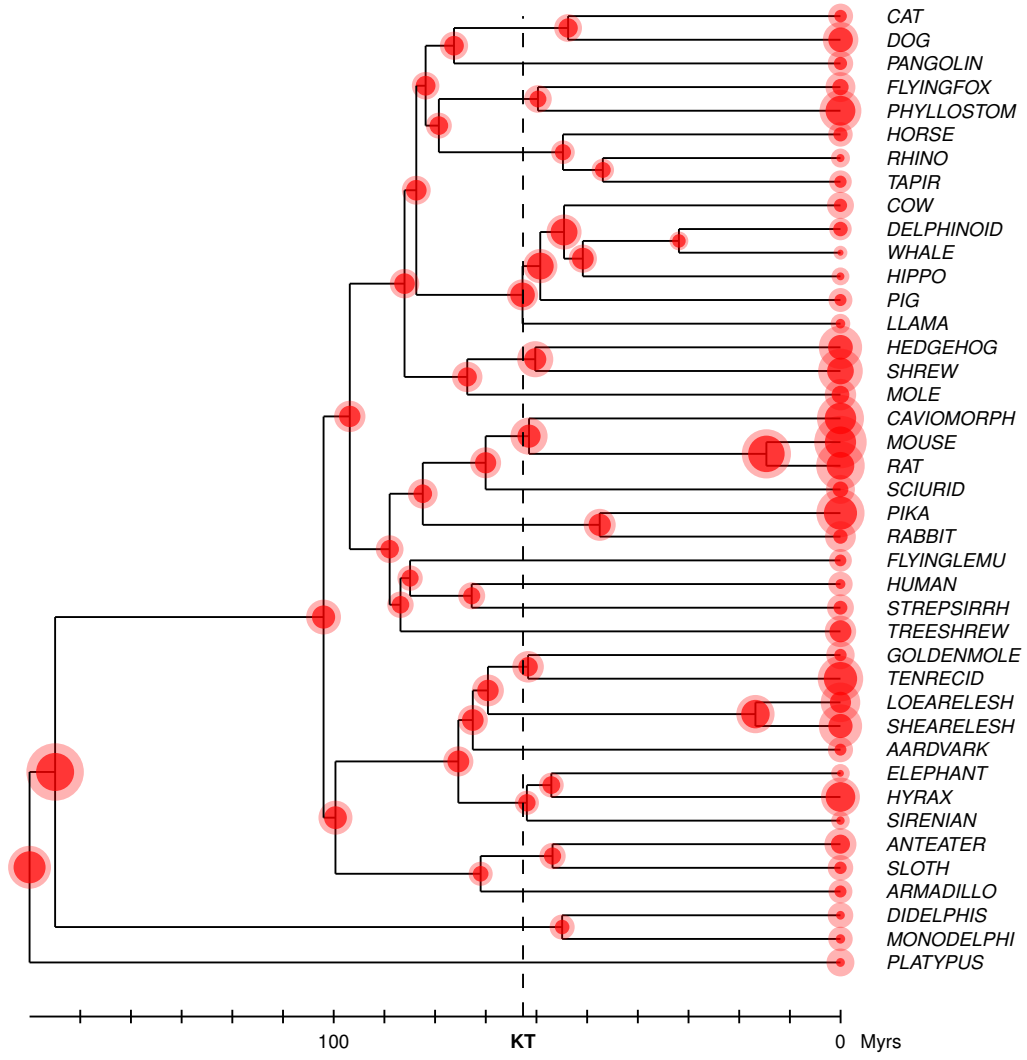
- diversification process (e.g. birth-death, parameters λ, μ, ρ)
- substitution rate: Brownian log-normal process (variance σ^2)
- substitution process (4x4 substitution matrix Q)
- complete model configuration: $\theta = (\lambda, \mu, \rho, \sigma, t, r)$

posterior distribution proportional to joint probability:

$$p(\lambda)p(\mu)p(\rho)p(\sigma) \quad p(\Psi \mid \lambda, \mu, \rho) \quad p(r \mid r_0, \Psi, \sigma)p(r_0) \quad p(D \mid r, \Psi, Q)$$

Posterior mean times and rates

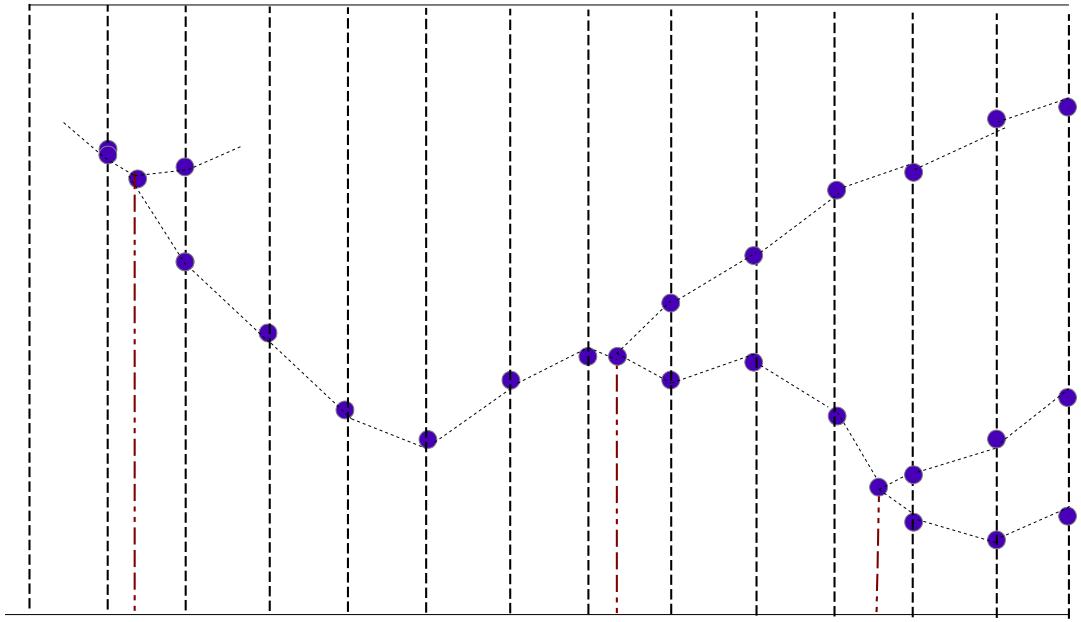




(Thorne et al 1998, Lepage et al 2007, Rannala and Yang 2007)

Multidivtime, PhyloBayes, MCMCtree, Beast

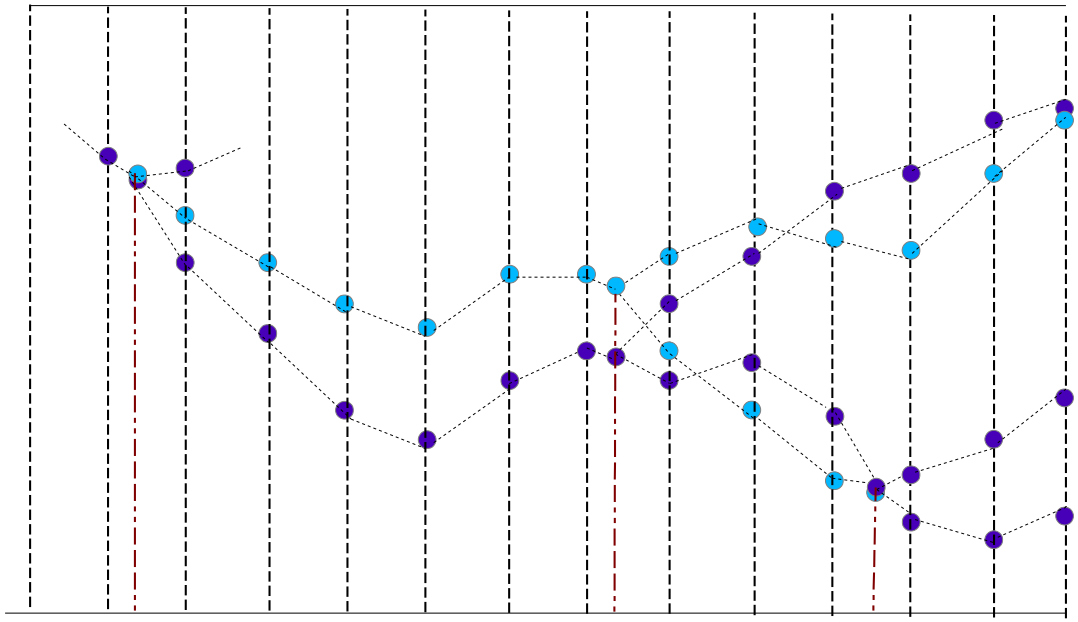
Getting away with branch-wise models



Fine-grained discretization schemes

- explicitly sample instant values of substitution rate
- special MCMC methods for resampling discretized Brownian paths
- can be used for modeling variation in any aspect of subst. process
- Horvilleur and Lartillot, 2014, Bioinformatics (in press)

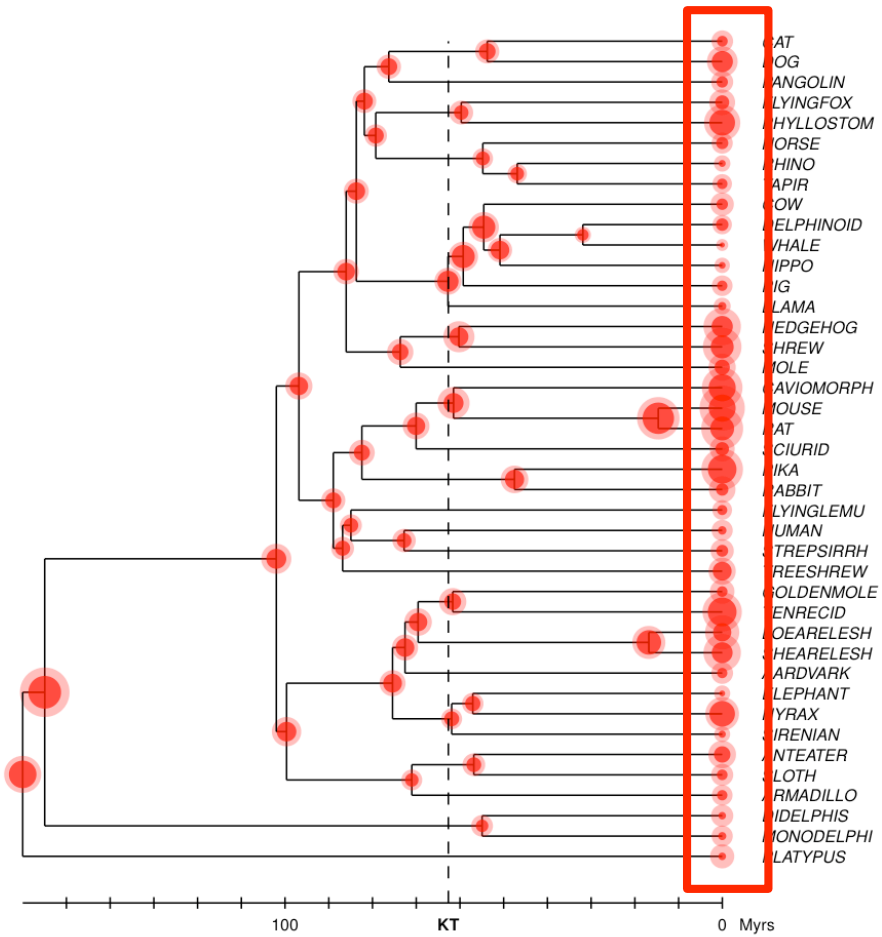
Getting away with branch-wise models

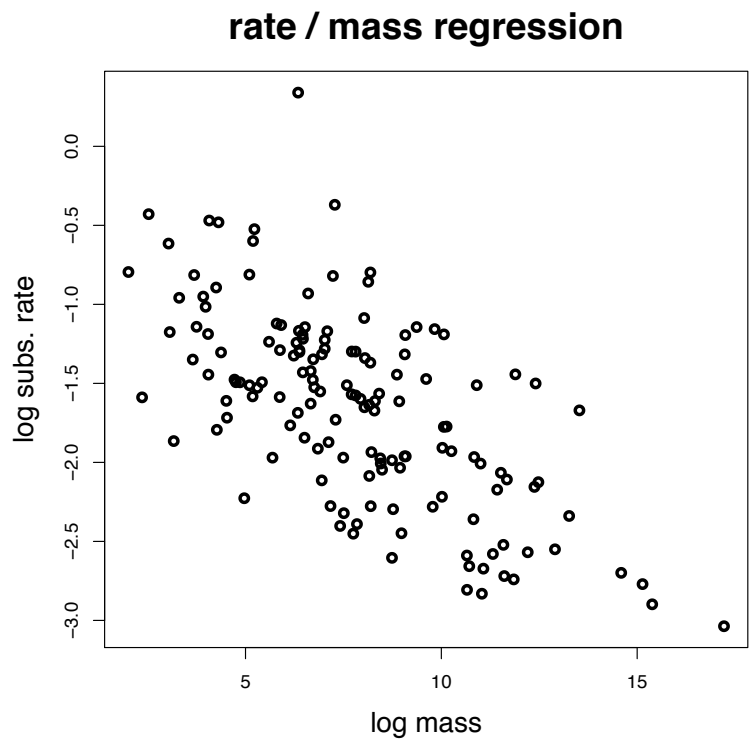


Fine-grained discretization schemes

- explicitly sample instant values of substitution rate
- special MCMC methods for resampling discretized Brownian paths
- can be used for modeling variation in any aspect of subst. process
- Horvilleur and Lartillot, 2014, Bioinformatics (in press)

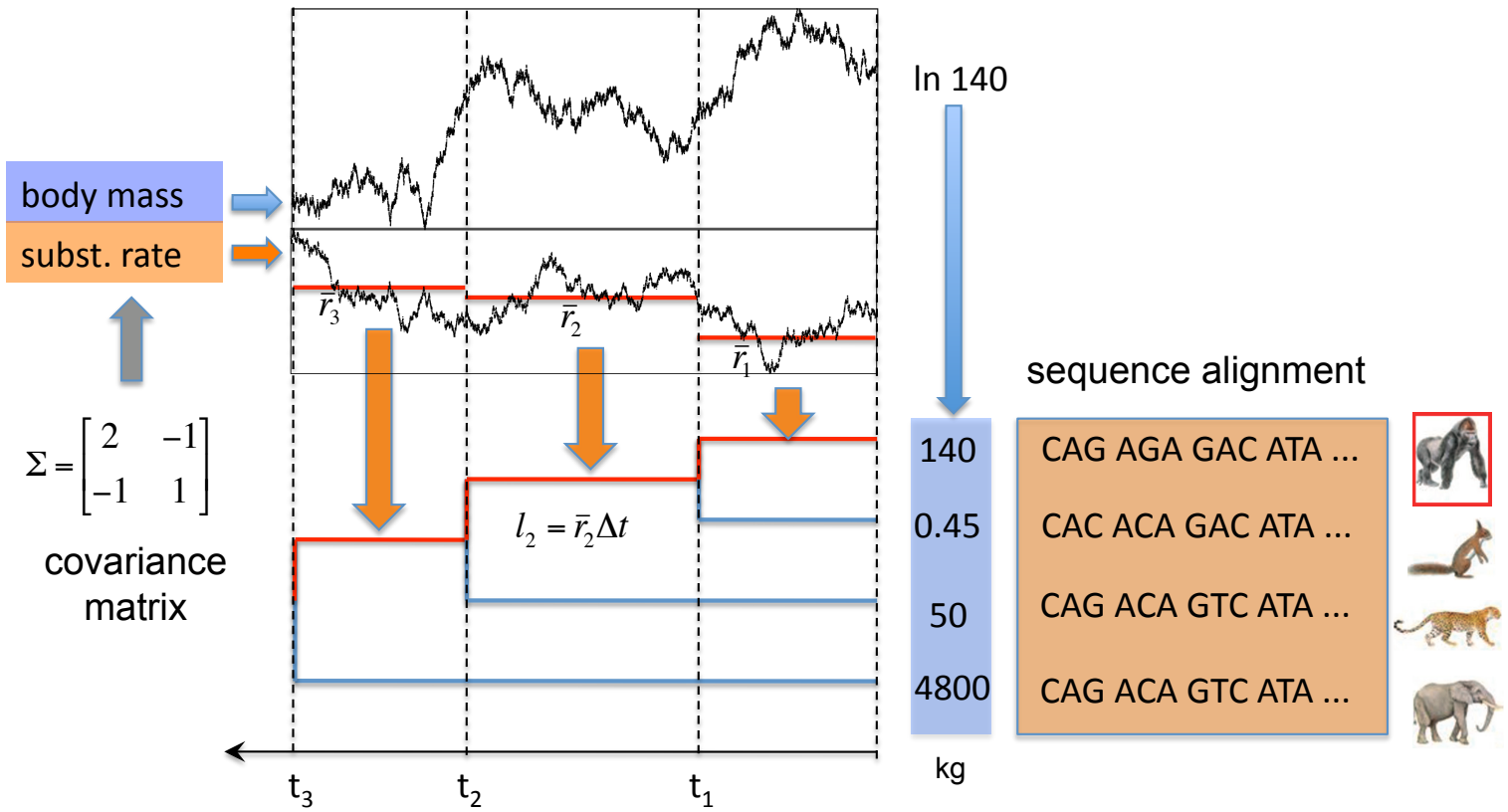
Correlating rates and traits





- sequential method: error propagation problems
- circularity in the way phylogenetic inertia is dealt with
- suggests a more direct *integrative* approach

The molecular comparative method

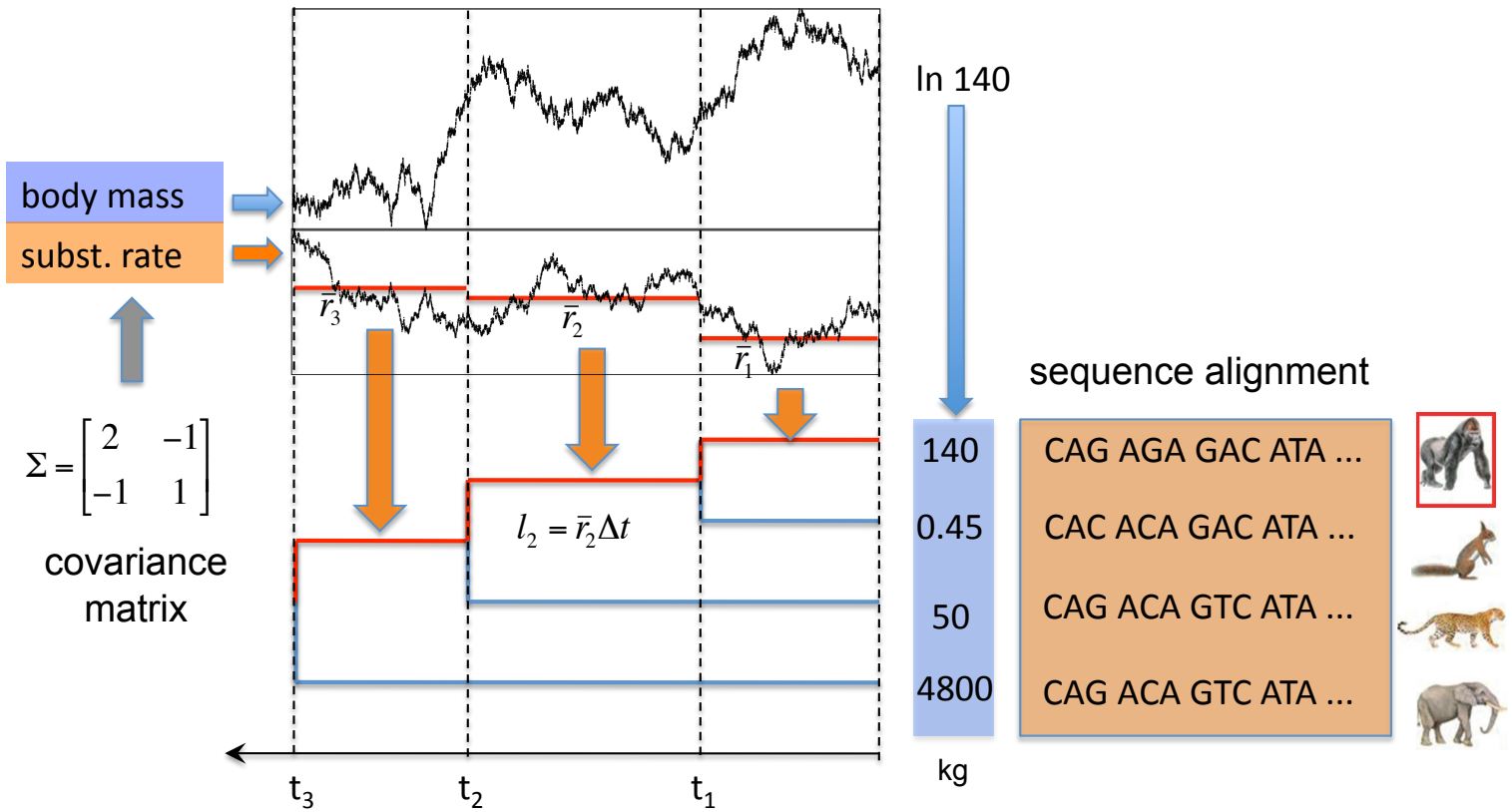


Lartillot and Poujol, 2011, Mol Biol Evol, 28:729

Hierarchial Bayesian model (parameter estimation by MCMC)

- diversification process t (birth-death, parameters λ, μ, ρ)
- Brownian multivariate process X (covariance matrix Σ)
- time-dependent codon model Q

The molecular comparative method



(Lartillot and Poujol, 2011, Molecular Biology and Evolution)

posterior proportional to joint probability:

$$p(\lambda, \mu, \rho)\, p(t \mid \lambda, \mu, \rho)\, p(\Sigma)\, p(X \mid t, \Sigma)\, p(D \mid X, t)$$

Generalization

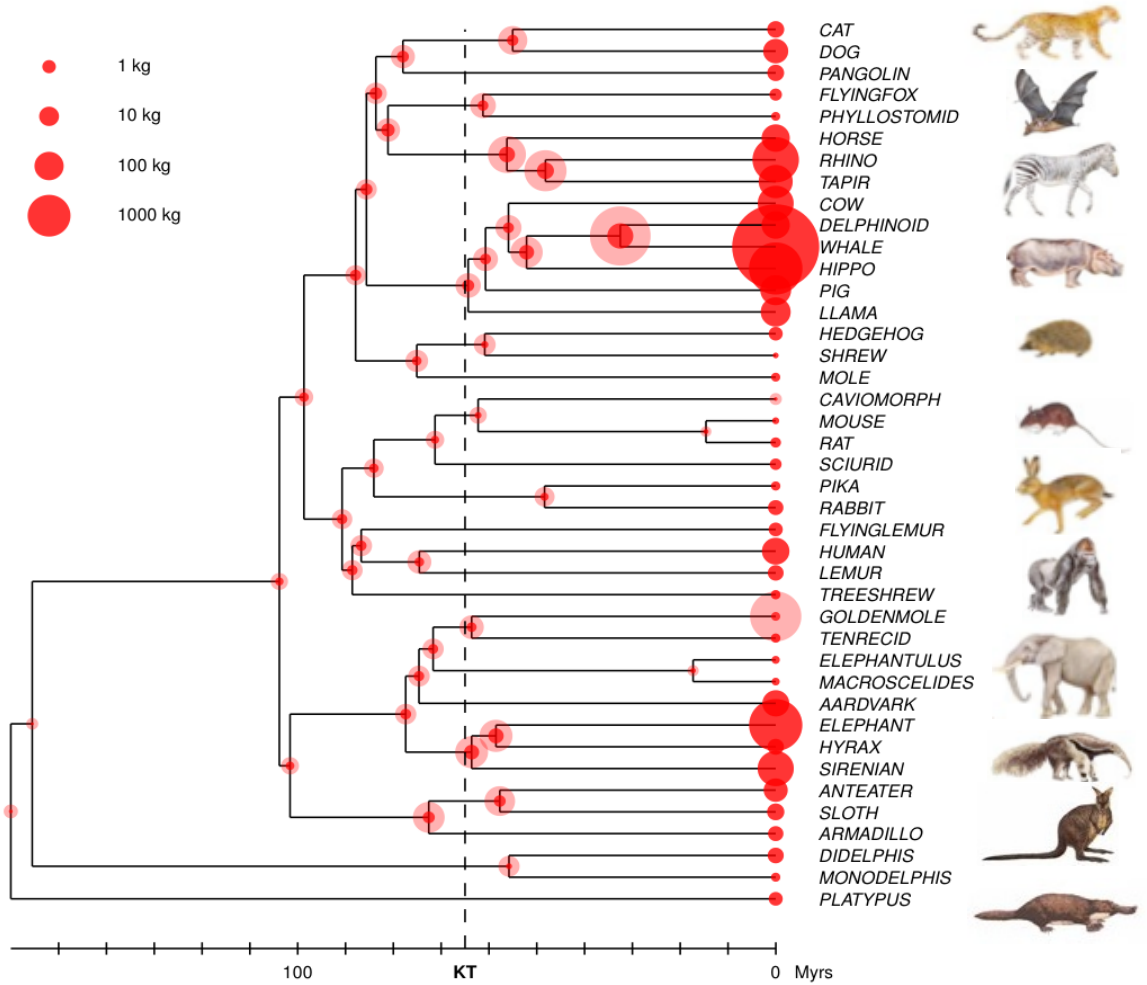
time-dependent substitution parameters

- rate of synonymous substitution (r)
- non-synonymous / synonymous ratio (ω)
- equilibrium GC composition (γ)

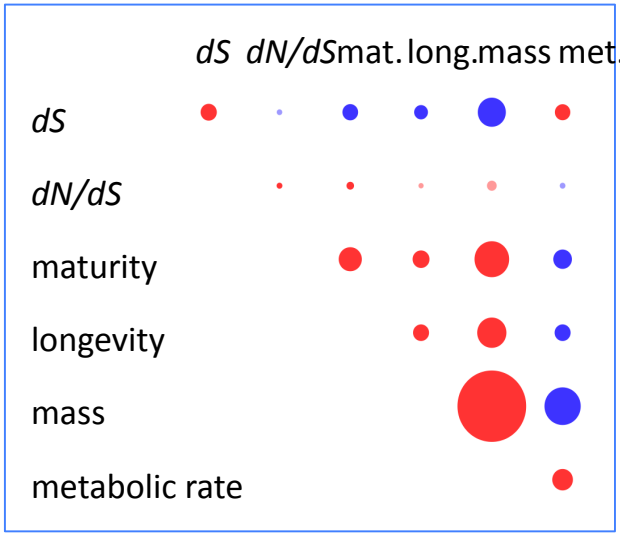
time-dependent quantitative traits

- sexual maturity (proxy of generation time)
- adult body mass
- maximum recorded lifespan (proxy of longevity)
- metabolic rate
- genome size
- etc.

Joint inference of rates, dates and traits



Alignment of 13 genes
4800 coding positions



red: positive correlation
blue: negative correlation

Lartillot and Delsuc, 2012, Evolution 66:1773