ESTIMATING EVOLUTIONARY RATES AND HISTORICAL DATES

Philippe Lemey[1] & Marc Suchard[2]

[1] Rega Institute, Department of Microbiology and Immunology, K.U. Leuven, Belgium.

[2] Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA. Department of Biostatistics, UCLA School of Public Health

Dates of historical events

MOLECULAR SEQUENCES

Alignment Methods — BIOINFORMATICS

ALIGNMENT

Sequence Evolution Models
Phylogenetic Methods — PHYLOGENETICS

EVOLUTIONARY TREE
(time scale = genetic distance)

Molecular Clock Models — PHYLOGENETICS

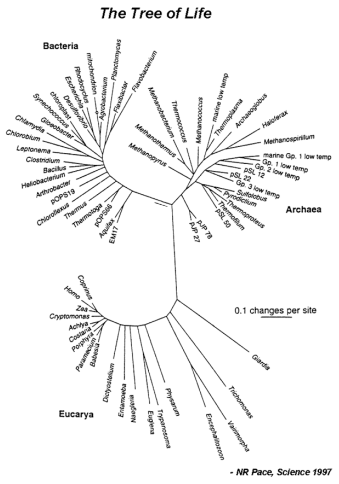
EVOLUTIONARY TREE
(time scale = years)

Coalescent Models — POPULATION GENETICS

EPIDEMIOLOGY

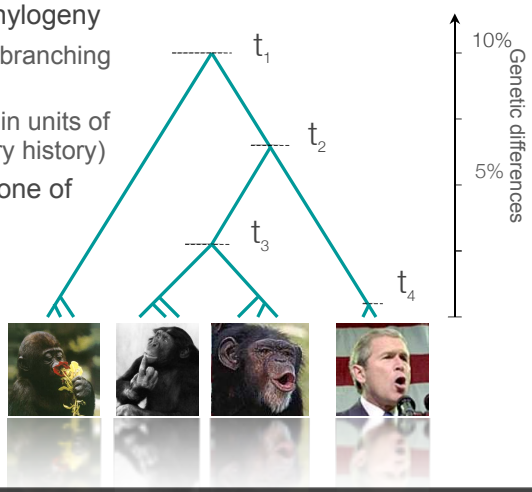
Molecular phylogenies

- most molecular phylogenies
  - are unrooted (or the rooting is due to prior information)
  - have branch lengths representing genetic change



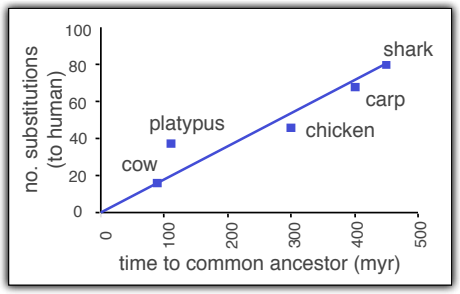
Molecular phylogenies

- the ideal molecular phylogeny
  - is rooted (implies a branching order)
  - has branch lengths in units of time (an evolutionary history)
- how do we construct one of these trees?



A constant evolutionary rate through time?

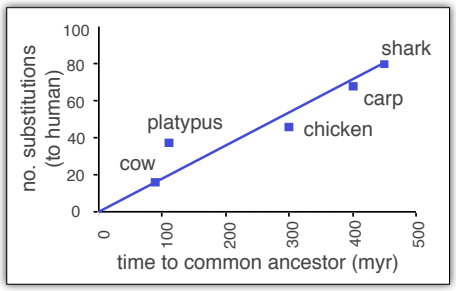
- to obtain a time phylogeny, the evolutionary model must assume a relationship between the accumulation of genetic diversity and time



- Zuckerkandl and Pauling (1962): the rate of amino acid replacements in animal haemoglobins was roughly proportional to real time, as judged against the fossil record

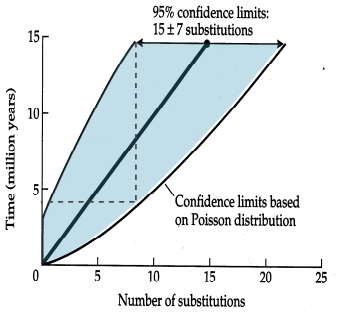
A constant evolutionary rate through time?

- the *molecular clock* is particularly striking when compared to the obvious differences in rates of morphological evolution...



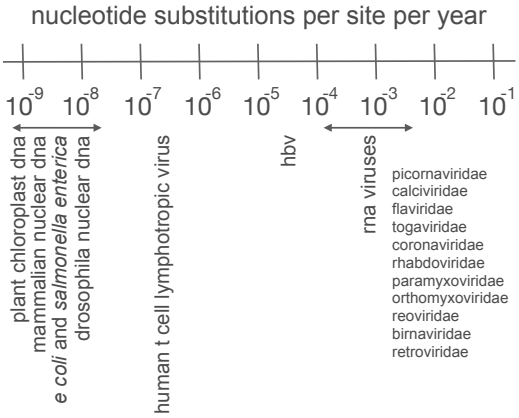
The molecular clock is not a metronomic clock

- if mutation every MY with Poisson variance
  - 95% of the lineages 15MY old have 8-22 substitutions
  - 8 substitutions also could be < 5 MY old



- Molecular Systematics, p532.

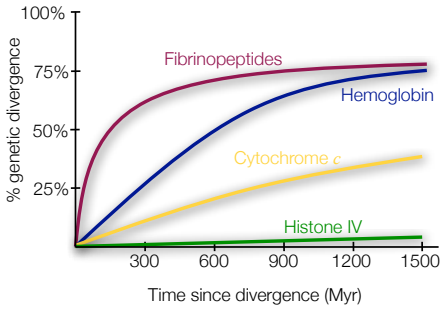
However, there is no global molecular clock



However, there is no global molecular clock

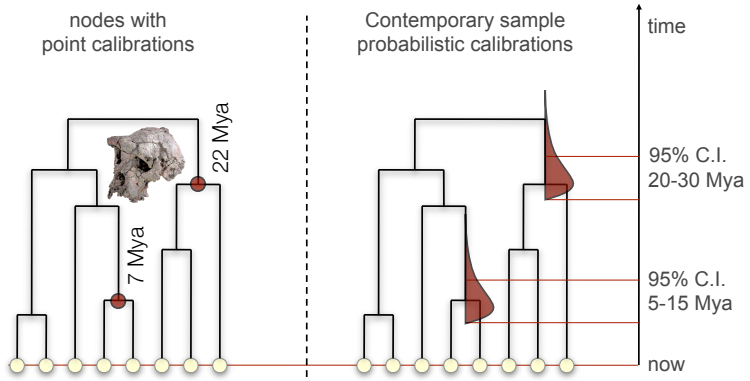
- different genes, different profiles
- variation in mutation rate?
- variation in selection

  genes coding for some molecules under very strong stabilizing selection



calibrating the molecular clock

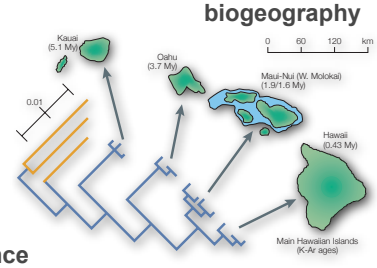
From substitutions units to time units



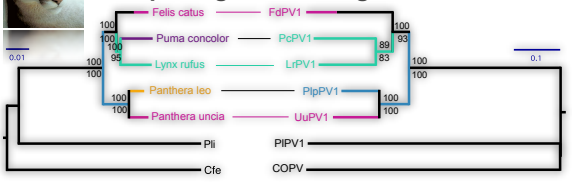
Node calibrations

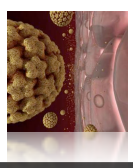
Fossils

biogeography



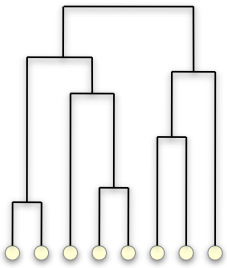
host-pathogen co-divergence



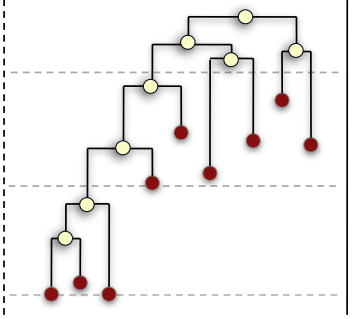


Tip calibrations

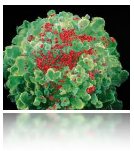
contemporary sample, no time structure



serial sample, with time structure



Tip calibration: 2 major sources



RNA viruses evolve quickly: 10^{-3} - 10^{-5} substitutions per site per year.

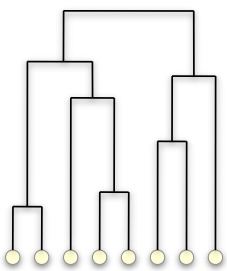
- Substitutions accumulate between the times of sampling
- Serially sampled sequences or heterochronous sequences

Measurably evolving population

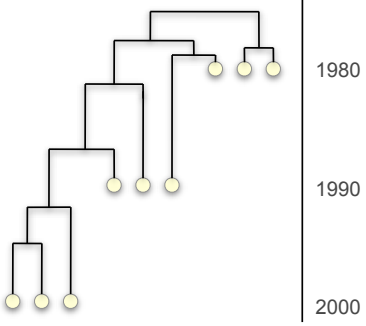
ancient DNA data sets of radiocarbon-dated specimens

Time structure via tip calibration

Contemporary sample no time structure



Serial sample with time structure



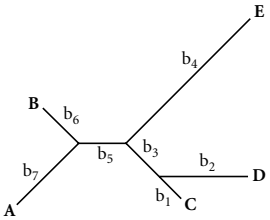
- Rambaut A. (2000) *Bioinformatics*, **16**, 395-399.

Molecular clock vs. non-clock

- strict molecular clock:
  Zuckerkandl & Pauling (1962) in Horizons in Biochemistry, pp. 189–225
  - all lineages evolve at the same rate
  - allows the estimation of the root of the tree and dates of individual nodes
- unconstrained (unrooted) Felsenstein model:
  Felsenstein (1981) *JME*, **17**: 368 - 376
  - each branch has its own rate independent of all others
  - time and rate are confounded and can only be estimated as a compound parameter (branch lengths)

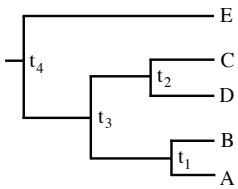
Likelihood ratio test of the molecular clock

- complex model H_1



2N-3 parameters

- null model H_0



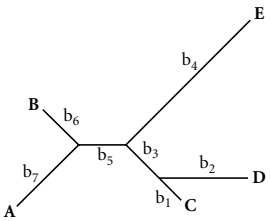
N-1 parameters

$LR = 2(\log L(H_a D) - \log L(H_0 D))$

- likelihood ratio test with N-2 degrees of freedom
- models are nested because values of b_1-b_7 can be specified that give node heights t_1-t_4

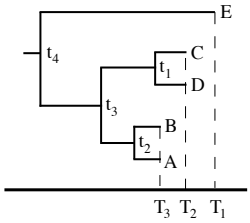
Likelihood ratio test of the *tip-dated* molecular clock

- complex model H_1



2N-3 parameters

- null model H_0



N parameters

- likelihood ratio test with N-3 degrees of freedom
- models are nested because values of b_1-b_7 can be specified that give node heights t_1-t_4 and evolutionary rate μ

Relaxing the molecular clock

Need for a relaxed molecular clock

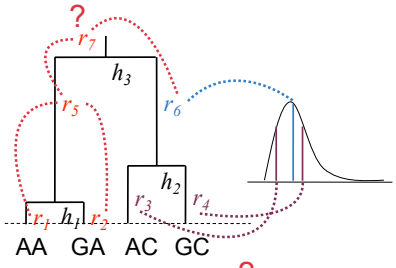
- the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum.
- dominate phylogenetic inference
- but both are biologically unrealistic:
  - the real evolutionary process lies between these two extremes
  - model misspecification can produce positively misleading results

Relaxed molecular clock methods

- Some phylogenetic methods allow the rate to vary among branches in a controlled manner
  - Local clock models (PAML, QDate)
  - Non-parametric rate smoothing (r8s)
  - Ad hoc heuristic rate smoothing (PAML)
  - Penalized likelihood (r8s)
  - Bayesian relaxed-clock methods (multidivtime, PhyBayes, BEAST)

Autocorrelated relaxed clocks

- rates for each branch are drawn from a distribution centered on the rate of the ancestor
  - but what is the rate at the root?
  - A prior degree of autocorrelation?
  - not currently possible to do phylogenetic inference

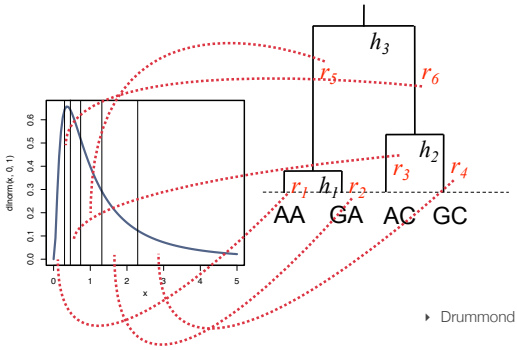


$$r_i \sim LogNormal(r_{A(i)}, \overset{?}{\sigma^2}\Delta t_i)$$

- e.g., Thorne JL, Kishino H, Painter IS (1998) Mol Biol & Evol **15**: 1647-1657.

Uncorrelated relaxed clocks

- rates for each branch are drawn independently from an identical distribution:



$$r \sim Exp(\lambda)$$

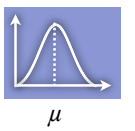
$$r \sim LogNormal(\mu,\sigma^2)$$

$$r \sim Gamma(\alpha,\beta)$$

- Drummond et al. (2006) Plos Biology **4**: e88.

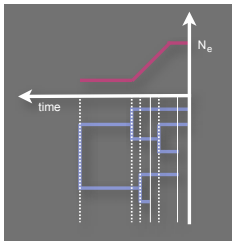
Bayesian evolutionary analysis sampling trees

- Given sequence data that is temporally spaced estimate true values of:
  - substitution parameters (μ and Q)
  - ancestral genealogy (g = E_g , t_γ)
    - tree topology
    - dates of divergence
  - population history (θ)



μ

Q



- Bayesian inference

$$P(g,\mu,\theta,Q|D) = \frac{1}{Z} Pr\{D|g,\mu,Q\} f_g(g|\theta) f_\mu(\mu) f_\theta(\theta) f_Q(Q)$$

"relaxed phylogenetics and dating with confidence"

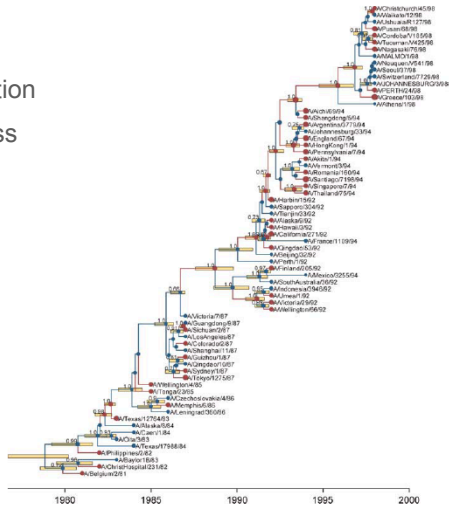
$t = \{ t_1, t_2, \ldots, t_{2n-1} \}$

$R = \{ r_1, r_2, \ldots, r_{2n-1} \}$

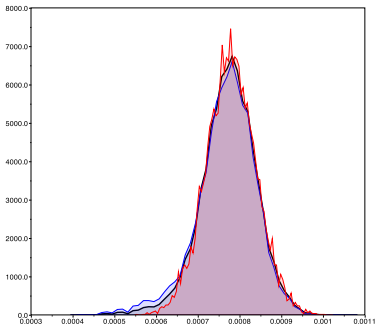
$$f(R|g) = f(R) = \prod_{i=1} \lambda e^{-\lambda r_i}$$

Uncorrelated relaxed clocks: example

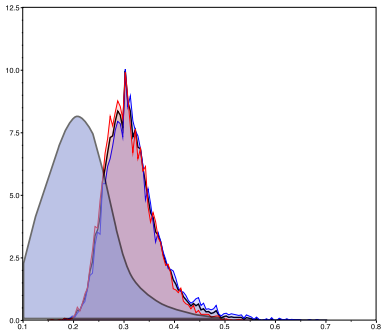
- Phylogenetic inference
- measuring autocorrelation
- measuring clocklikeness



Evaluating clock-like behaviour?



rate



stdev

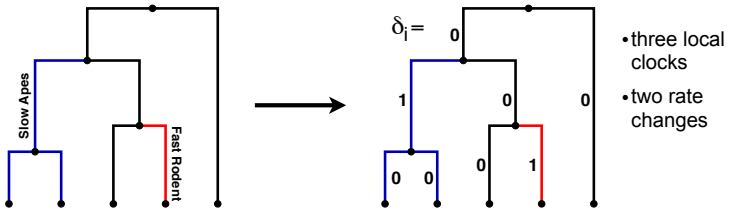
Model testing using Bayes factors

- posterior $p(\theta|D,M) = \dfrac{p(D|\theta,M)\,p(\theta|M)}{p(D|M)}$
- marginal likelihood $p(D|M) = \int_{\theta} p(D|\theta,M)\,p(\theta|M)\,d\theta$
- Bayes factor $B_{01} = \dfrac{p(D|M_1)}{p(D|M_2)}$
  - Harmonic mean estimator
    Newton and Raftery, 1994; Suchard et al., 2003
  - Path sampling (thermodynamic integration)
    Gelman, 1998; Ogata, 1989; Lartillot and Philippe, 2006

now in BEAST!
(Baele et al., MBE, 2012)

Random local clocks

- Rate changes do not necessarily occur regularly or on every branch
- Small number of significant changes
- Can we handle the uncertainty in the number and locations of (a small number of) local clocks?

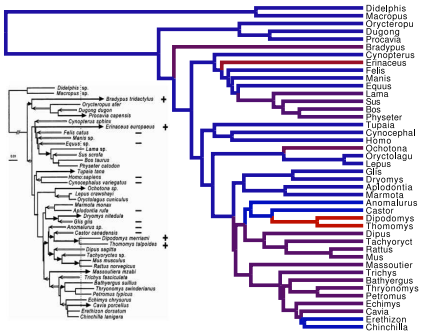


- three local clocks
- two rate changes

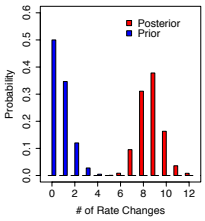
- How to explore 2^{2n-2} clock models?

Random local clocks

- Using Bayesian stochastic search variable selection: formulate a prior that such that many rate changes (indicators) are 0 but allow the data to determine which ones are required to explain (most of the) rate variation using MCMC



- Three mtDNA nuclear genes from 42 mammals (Douzery, 2003)
- 5-12 local clocks



Drummond and Suchard, 2010.