13. Molecular clocks

EEOB563

Spring 2021

> "Molecular clock is best understood as a statement about the ability to infer time from genetic sequence data, not as a claim about a regular 'tick rate' of changes in the genome". Bromham et al., 2008.

1 The beginning

In the beginning, our knowledge of evolutionary time was based exclusively on the fossil record. In 1962 and 1965, Zuckerkandl and Pauling proposed that for any given lineage, the rate of molecular evolution (amino acid substitutions per year) is constant over time and estimated this rate for haemoglobin molecules in mammals at 1/107 years/100aa. They also coined the term "molecular evolutionary clock" [link]

This finding was unexpected for two reasons:

1. it had been assumed that, as with morphological evolution, there would be large variation in the rate of change both between species and over evolutionary time
2. although the rate seems low, too many substitutions would be expected in the genome

Note further: the rate appears to be proportional to time, and not to the number of generations or cell divisions. The independence of generation time speaks against positive selection as a driving force of evolution. The independence of cell cycles suggests that most mutations do not happen during replication.

2 The rate of molecular evolution under a molecular clock

If DNA evolves in a neutral fashion, then its rate of evolution will be equal to the mutation rate, μ. The mean estimate of the amount of change separating two sequences will equal $d = 2\mu t$.

However, "ticks" of the clock are stochastic, not deterministic. The simplest way to model accumulation of mutations over time is by using a Poisson process. Accordingly, the probability that i mutations will occur during time t is given by:

$$p(i) = PX = i = e^{-\lambda}\frac{\lambda^i}{i!} \tag{1}$$

with the expected value and variance are equal to $\lambda = \mu t$ (Fig. 1)

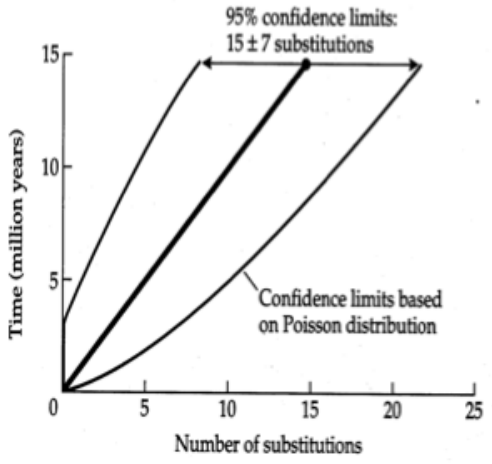

Figure 1: The sampling expectations of a perfect molecular clock, arbitrary set at one substitution per sequence, per million years. From Hillis, Mable and Moritz, 1996

If there is one mutation every MY, after 15MY, 95% of lineages will have 8-22 substitutions. However, 8 substitutions could also be accumulated in 5 MY.

Note that the models of sequence evolution that we have discussed in class are Poisson models.

Also note that since the mean and variance of the clock are equal under the Poisson model, we can calculate the index of dispersion ($R(t)$) as the ration of $Var(\#$ of mutations) to $E(\#$ mutations)

In many cases $R(t) > 1$ or the clock is overdispersed. Rate of molecular evolution can differ between

- nucleotide positions
- genes
- genomic regions
- genomes (nuclear vs organelle),

Rates can only differ for the same gene in different lineages due to

- Changes in mutation rate due to
  - mutation rate
  - repair efficiency
  - generation time
- Changes in population size
  - Nearly neutral theory of molecular evolution

- Changes in selective coefficient

Thus, it is generally acknowledged that strict molecular clock can't be applied globally or to distantly related species. However, for closely related species, or in the analysis of population data, the molecular clock is a good approximation of reality.

3 Beyond the strict molecular clock

There are several ways to deal with violations of molecular clock assumptions:

1. Use strict clock but filter the data
    - Gene selection
    - Lineage selection
    - Note, that both approaches reduce the amount of data (increase variance).
2. Local clocks (based on the assumption that closely related lineages share similar rates of evolution)
    - *Ad hoc* local clock (partitioning is guided by a priory biological information
    - Dirichlet Process Prior (DPP) local clocks (the number of rate categories and assignment of branches to them are treated as random variables under the DPP model).
3. Relaxed clocks
    - Autocorrelated models (assume that neighbouring branches share similar rates). Rate autocorrelation is modeled by drawing the rate along each branch from a lognormal, gamma, or exponential distribution with mean equal to the rate along the parent branch.
    - Uncorrelated models (no assumption of correlation between rates in neighboring branches). Branch-specific rates are sampled from a single distribution (lognormal, exponential, or gamma).

4 Calibration issues

In order to get absolute rather than relative estimates of divergences, molecular clock needs to be calibrated. Although the issue of calibration is often glossed over in molecular clock studies, uncertain calibrations can be the main source of uncertainty in such studies.

Evidence from the fossil record is most commonly used for the purpose of calibration. Initially fossil-age calibrations were used as point values for clades. In Bayesian clock dating, calibration information is incorporated through the prior on times. Fossil ages provide good minimum-age bounds on clade ages, but they are insufficient for calibrating a molecular tree. Instead soft bounds and arbitrary curves are used as calibrations. The issue of fossil calibration has been recently reviewed and several best practices suggested (Parham et al. Syst. Biol. 61(2):346–359, 2012):

1. Museum numbers of specimen(s) that demonstrate all the relevant characters and provenance data should be listed. Referrals of additional specimens to the focal taxon should be justified.
2. An apomorphy-based diagnosis of the specimen(s) or an explicit, up-to-date, phylogenetic analysis that includes the specimen(s) should be referenced.
3. Explicit statements on the reconciliation of morphological and molecular data sets should be given.
4. The locality and stratigraphic level (to the best of current knowledge) from which the calibrating fossil(s) was/were collected should be specified.
5. Reference to a published radioisotopic age and/or numeric timescale and details of numeric age selection should be given.

There is an additional issue of node calibration vs. tip calibration, which was recently reviewed in O'Reilly et al. (2015).

You can find much more information in these excellent recent reviews:

Bromham et al. 2018. Bayesian molecular dating: opening up the black box. Biol. Rev. 93, 1165–1191. [link]

dos Reis et al. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nature Reviews Genetics 17, 71–80. [link].

Magallón 2020. Principles of Molecular Dating. The Molecular Evolutionary Clock. [link]