Encyclopedia of Scientific Dating Methods
DOI 10.1007/978-94-007-6326-5_18-1

Molecular Dating of Evolutionary Events

Q1 David Duchene* and Lindell Bromham
Centre for Macroevolution and Macroecology, Division of Evolution, Ecology, and Genetics, Research School of Biology, Australian National University, Canberra, ACT, Australia

Definition

Inference of the age of evolutionary events using statistical analysis of rates of change of DNA or amino acid sequences.

Introduction

Molecular dating is used in the biological sciences to estimate the age of evolutionary events. Changes to DNA and amino acid sequences accumulate continuously in the genome over time, so comparing DNA sequences between lineages allows us to estimate the time since they last shared a common ancestor. However, the rate of change varies across the genome and among species. So in order to use molecular data to date evolutionary events, we need a way of estimating the rate of change in genetic sequences over time for any given dataset. Molecular dating requires a set of homologous genetic sequences (all related by descent from a single ancestral sequence), a method for inferring the number of changes that have occurred during the evolution of these sequences, and calibrating information to estimate their rate of change.

An increasing variety of analytical tools are available for molecular dating of evolutionary events. All assume that the rate of genetic sequence evolution has some degree of predictability, so that the age of lineage divergence can be estimated from molecular data using a model of evolutionary change and some information about timing of evolutionary events (typically derived from fossils, biogeography, pedigree, or ancestral genetic sequences).

Methods for Dating Evolutionary Events

Molecular dating methods begin by estimating the amount of genetic difference that has occurred between sequences. Then, assumptions about the rate of change are used to infer the amount of time needed to explain that amount of genetic difference. The difference between genetic sequences is estimated using a model of the frequency of each type of substitution. Models of molecular evolution form the basis of methods to estimate the relatedness of multiple sequences and the timing since their evolutionary divergence (for more detail on the models, see Yang 2006). These models can be simple and give an equal probability to every type of substitution (Jukes and Cantor 1969), or be complex and account for the attributes of specific genetic datasets (e.g., Yang 1994). For instance, the mitochondrial ND6 genetic region in marine mammals has virtually no transversions (changes from the purines, A and G, to pyrimidines, C and T, or vice versa), so it

*Email: david.duchene@anu.edu.au

DOI 10.1007/978-94-007-6326-5_18-1

is appropriate to use parameter settings that model a large number of transitions (changes from purine to another purine, or pyrimidine to pyrimidine) (Duchêne et al. 2011).

The simplest way to use genetic difference estimates to infer the timing of evolutionary divergences is to select "clocklike" data, where the amount of genetic difference accumulates at the same rate in all lineages. There are methods to detect significant variation in rates (Langley and Fitch 1973; Takezaki et al. 1995). For example, Likelihood Ratio Test can be used to ask whether data can be best described by a single uniform rate, or a multiple rate model (Brown and Yang 2011). Importantly, while some datasets may approximate uniform rates, clocklike behavior cannot be generalized to a particular gene, lineage, or taxonomic level.

Another way to account for rate variation is to have multiple rate categories in the same analysis. For example, sections of the dataset may have their own "local clock" (Drummond and Suchard 2010; Rambaut and Bromham 1998; Yoder and Yang 2000). Other methods account for rate variation by "relaxing" the molecular clock constraint on all branches, so every branch in a phylogeny has a different rate (e.g., Drummond et al. 2006; Yang 2007); for more on this topic, see Springer Reference article Relaxed Molecular Clocks; for reviews on dating methods, (see Magallón 2004; Rutschmann 2006; Welch and Bromham 2005).

Calibrating the Rate of Substitutions

Some molecular dating analyses use an assumed rate of substitutions to estimate dates. However, given that rates of molecular evolution vary across the genome and between lineages, it is preferable to use independent information to calibrate rates of molecular evolution for each dataset analyzed. The most common way of calibrating rates is to use a known date of one or more divergence events in the phylogeny.

The calibration must correspond to the earliest possible age of a node in the phylogeny. The dates of divergence events are rarely known with certainty, and the confidence limits on calibrations can be very large. In the case of fossil evidence, part of the uncertainty comes from the determination of the age of the fossil itself, drawn from stratigraphy and isotopic composition of the fossil (Benton and Donoghue 2007). However, the exact relationship between the fossil taxon and the divergence in the phylogeny is also typically unknown (Sauquet et al. 2012). A fossil taxon is unlikely to be identifiable as a member of a particular lineage until some time after the origin of the lineage, when key diagnostic characters have had time to evolve. So fossil dates typically represent minimum ages of lineages: they provide evidence that a lineage must have originated some time before that date (Bromham et al. 1999).

One way to account for calibration uncertainty is to have a probability distribution of node ages. A hard-bound can define the latest or the earliest point where a divergence may have occurred (also called the minimum age and the maximum age, respectively). While a minimum age can be defined by the age of a fossil known to occur on the lineage in question, because the lineages must have originated before the occurrence of the first fossil, special caution is required when placing a maximum constraint because it assumes with total confidence that a lineage was absent before a given time (Benton and Donoghue 2007; Ho and Phillips 2009; Hug and Roger 2007).

To overcome the strict assumptions of hard-bounded calibrations, an alternative is to use soft-bounded calibrations, so a nonzero probability can be assigned to all ages (Yang and Rannala 2006). To define a soft-bounded calibration, the user must specify a probability distribution for the date of the calibrating divergence. Given that the confidence in particular date estimates is specific to current knowledge, expert advice is often required to define these calibrations. Broadly,

DOI 10.1007/978-94-007-6326-5_18-1

a normally distributed calibration is conservative for fossil data that is imprecise or equivocal. A normally distributed calibration can also be used for biogeographic data that provides reasonable evidence for causing a divergence event, or for calibrations based on previous date estimation analyses (also called secondary calibrations, Ho and Phillips 2009). Lognormal distributions are often appropriate for fossil data as they provide a hard minimum bound, which is the age of the fossil. They also allow a higher probability for the divergence of a node at a point slightly older than the age of the fossil. These distributions are also applied when using previous date estimates from Bayesian analyses that produced lognormal distributions.

Exponentially distributed calibrations can be used if there is little evidence that an event occurred at any point before the fossil, or if the fossil is likely to be very close to the divergence event (Ho and Phillips 2009). To choose the values of the parameters of these distributions, the user must be well informed about the sources of error in the calibration. Poor calibrations may be useful if their uncertainty can be quantified, but if only poor calibrations are used, then the uncertainty in date estimates will be significant.

Summary and Conclusions

The field of molecular dating of evolutionary events can be divided by the methods to deal with variation in the rate of molecular substitutions. To date evolutionary events using molecular data, it is necessary to evaluate and choose a method to model molecular evolution. It is also necessary to find information to calibrate rates of substitutions, and all analyses must account for calibration uncertainty. Although modern software for this purpose can be deceptively simple, there is a plethora of considerations required before performing molecular dating analyses. In particular, Bayesian methods for molecular dating can involve complex assumptions about the model of molecular evolution and arguably are not entirely understood. Methods in place to compare and choose the most appropriate dating schemes are computationally demanding and may be severely biased (Baele et al. 2012; Lartillot and Philippe 2006; Xie et al. 2011). However, with appropriate information and a clear understanding of the uncertainty inherent in molecular date estimates, molecular dating is a powerful tool that can be extended to study the realms of evolution, ecology, biogeography, and past population processes.

Bibliography

Baele, G., Lemey, P., Bedford, T., et al., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, **29**, 2157–2167, doi:10.1093/molbev/mss084.

Benton, M. J., and Donoghue, P. C. J., 2007. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution*, **24**, 26–53, doi:10.1093/molbev/msl150.

Bromham, L., Phillips, M., and Penny, D., 1999. Growing up with dinosaurs: molecular dates and the mammalian radiation. *Trends in Ecology & Evolution*, **14**, 113–118.

Brown, R. P., and Yang, Z., 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, **11**, 271, doi:10.1186/1471-2148-11-271.

Drummond, A. J., and Suchard, M. A., 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Evolutionary Biology*, **8**, 114, doi:10.1186/1741-7007-8-114.

DOI 10.1007/978-94-007-6326-5_18-1

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88, doi:10.1371/journal.pbio.0040088.

Duchêne, S., Archer, F. I., Vilstrup, J., et al., 2011. Mitogenome phylogenetics: the impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. *PloS One*, **6**, e27138, doi:10.1371/journal.pone.0027138.

Ho, S. Y. W., and Phillips, M. J., 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology*, **58**, 367–380, doi:10.1093/sysbio/syp035.

Ho, S. Y. W., Phillips, M. J., Drummond, A. J., and Cooper, A., 2005. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular Biology and Evolution*, **22**, 1355–1363, doi:10.1093/molbev/msi125.

Hug, L. A., and Roger, A. J., 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. *Molecular Biology and Evolution*, **24**, 1889–1897, doi:10.1093/molbev/msm115.

Jukes, T. H., and Cantor, C. R., 1969. Evolution of protein molecules. In Munro, H. (ed.), *Mammalian Protein Metabolism*. New York: Academic, pp. 21–132.

Langley, C. H., and Fitch, W. M., 1973. *The Constancy of Evolution: A Statistical Analysis of a and b Haemoglobins, Cytochrome c, and Fibrinopeptide A. Genetic Structure of Populations*. Honolulu, HI: University of Hawaii Press, pp. 246–262.

Lartillot, N., and Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, **55**, 195–207, doi:10.1080/10635150500433722.

Magallón, S., 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. *International Journal of Plant Sciences*, **165**, S7–S21.

Rambaut, A., and Bromham, L., 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution*, **15**, 442–448.

Rutschmann, F., 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Diversity and Distributions*, **12**, 35–48, doi:10.1111/j.1366-9516.2006.00210.x.

Sauquet, H., Ho, S. Y. W., Gandolfo, M. A., et al., 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). *Systematic Biology*, **61**, 289–313, doi:10.1093/sysbio/syr116.

Takezaki, N., Rzhetsky, A., and Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution*, **12**, 823–833.

Welch, J. J., and Bromham, L., 2005. Molecular dating when rates vary. *Trends in Ecology & Evolution*, **20**, 320–327, doi:10.1016/j.tree.2005.02.007.

Xie, W., Lewis, P. O., Fan, Y., et al., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, **60**, 150–160, doi:10.1093/sysbio/syq085.

Yang, Z., 1994. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, **39**, 105–111.

Yang, Z., 2006. *Computational Molecular Evolution*. Oxford: Oxford University Press.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591, doi:10.1093/molbev/msm088.

Yang, Z., and Rannala, B., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, **23**, 212–226, doi:10.1093/molbev/msj024.

Yoder, A. D., and Yang, Z., 2000. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, **17**, 1081–1090.