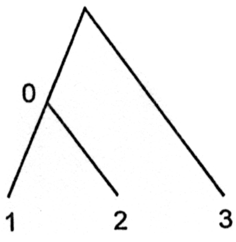
MOLECULAR CLOCK

The hypothesis of a molecular clock asserts that the rate of DNA or protein sequence evolution is constant over time or among evolutionary lineages. In the early 1960s, when protein sequences became available, it was observed that the rate of evolution for proteins, such as hemoglobin, were relatively constant among different orders of mammals. The observation led to the proposal of the molecular clock hypothesis by Emile Zuckerkandl and Linus Pauling in 1965. The proposal had an immediate impact on the development of the field of molecular evolution. First, the utility of the molecular clock was obvious from the beginning. If proteins evolved at constant rates, they can be used to reconstruct phylogenetic relationships among species and to estimate the dates of species divergences. Second, the accuracy of the clock and the mechanism of molecular evolution have been a focus of controversy. At the time, the synthetic theory of evolution or neo-Darwinism, which maintains that the rate of evolution is determined by environmental changes and natural selection, was generally accepted by evolutionists. A constant rate of evolution among species as different as elephants and mice was unthinkable. For example, morphological characteristics are well known to have markedly different rates of evolution among lineages. Motoo Kimura proposed the neutral theory of molecular evolution in the late 1960s, and the molecular clock was immediately solicited as a major piece of supporting evidence (Kimura, 1983). This theory maintains that most molecular evolution is dominated not by natural selection but by random fixation of neutral mutations, that is, mutations whose effects on fitness are too small for natural selection to play a role in determining their fate. The rate of molecular evolution is then equal to the neutral mutation rate, independent of factors such as environmental changes and population sizes. If the mutation rate was similar and the function of the protein remains the same among lineages so that the same proportions of mutations are neutral, a constant evolutionary rate is predicted by the neutral theory.

FIGURE 1. A Phylogeny of Three Species to Explain the Relative Rate Test.
Drawing by Ziheng Yang.



MAXIMUM LIKELIHOOD ESTIMATION AND LIKELIHOOD RATIO TEST

Maximum likelihood is a general methodology for estimating statistical parameters. Suppose the probability of observing the data (D) is $P(D; \vartheta)$, where ϑ are parameters under the model. Because the data are observed, we view P as a function of the unknown parameters and write it as $L(\vartheta; D) = P(D; \vartheta)$. L is known as the likelihood function. The values of ϑ that maximize L, or equivalently its logarithm, $l = \ln\{L\}$, are the maximum likelihood estimates.

The likelihood ratio test plays a central role in hypothesis testing. Suppose the more general (alternative) model has p parameters with log likelihood l_1, and the simpler (null) model has q parameters with log likelihood l_0. Then twice the log likelihood difference, $2\Delta l = 2(l_1 - l_0)$, is approximately χ^2 distributed with d.f. $= p_1 - p_0$, if H_0 is true. If the observed value of the test statistic $2\Delta l$ is greater than the χ^2 critical value, we reject H_0.

ZIHENG YANG

The concept of molecular clock has a long history of controversy. It is often a focus of hot debate even today. Early controversies were about whether the clock held, and when it did not, what factors might be responsible for the rate differences among lineages. For example, rodents were suggested to evolver faster than primates; one hypothesis proposed that rodents have a shorter generation time, more germ-line cell divisions per calendar year, and thus a higher mutation rate. Since the 1980s, DNA sequences have become widely available, and these reveal that the molecular clock is violated for most genes or species groups, except for sequences from very closely related species.

More recent controversies have been focused on the most common use of the clock assumption, that is, the dating of divergence times. Dating using molecules has produced a steady stream of controversies, as the molecular dates are often at odds with the fossil records or with the current interpretations of the fossil and morphological data. Two particular examples have attracted much attention, the Cambrian "explosion" about 545 million years ago and the mammalian "radiation" about 65 million years ago. In each case, molecular studies produced dates much older than indicated by the fossil data, sometimes twice as old. Part of the discrepancy probably arises from the incompleteness of fossil data; fossils represent the time when species developed diagnostic morphological characters and were fossilized, and molecules represent the time when the species stopped intermingling, so fossil dates have to be younger

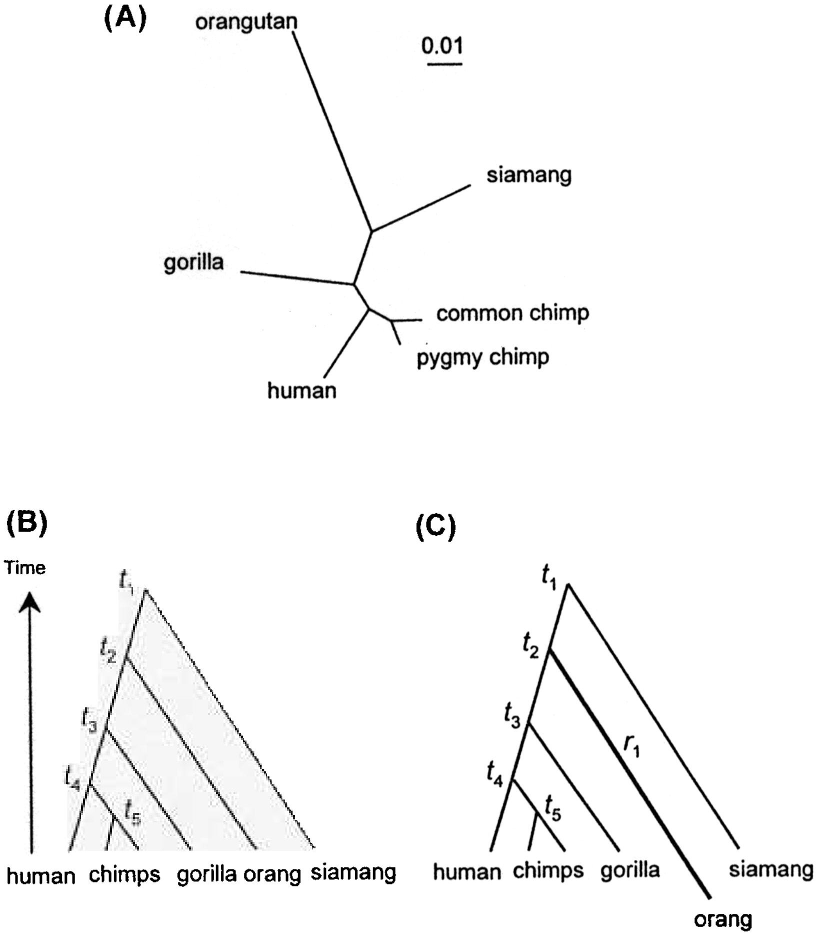


FIGURE 2. Phylogenies of Six Hominoid Species for Testing the Clock Assumption and Estimating Divergence Dates.
(A) The unrooted tree without assuming the clock. This is the alternative model, and the unknown parameters include the nine branch lengths. (B) Under the (global) clock assumption, the parameters are the five node distances, the distances of the five internal nodes from the present time, measured by the expected number of nucleotide substitutions per site. (C) A local clock model that assigns a different evolutionary rate to the orangutan lineage. From Yang, 1996.

than molecular dates. Part of the discrepancy seems to arise from inaccuracies in molecular date estimation. In particular, molecular date estimation is very sensitive to violation of the molecular clock (see below).

This article discusses statistical tests of the clock assumption, as well as methods for using the global and local clock models to estimate divergence dates.

A few clarifications are in order. First, the molecular clock was envisaged as stochastic. Molecular changes accumulate at random according to a statistical Poisson process, so that random fluctuations are expected, although the underlying rate is constant over time. Second, different genes or proteins, or different regions of the same gene, may have very different evolutionary rates, and their clocks tick at different rates. The interpretation offered by the neutral theory for this observation is that different genes are under different selective constraints, and those under stronger constraints will have a smaller proportion of neutral mutations and a lower evolutionary rate. Third, the molecular clock is not expected to be universal and is usually applied to a group of species. For example, we might say that the clock holds for a gene within mammals.

Tests of the Molecular Clock. A number of statistical tests have been designed to examine the molecular clock hypothesis. The simplest is the relative rate test based on estimates of pair-wise sequence distances. To test whether species 1 and 2 have the same rate, we get an outgroup species 3, which is more distantly related (see Figure 1). If the clock hypothesis is true, the distance (expected amount of change) from ancestral node 0 to species 1 will be the same as the distance from node 0 to species 2; that is, $d_{01} = d_{02}$. Because we do not have data for the ancestral node 0, we test the clock hypothesis by testing $d_{13} = d_{23}$, using the statistic $d = (d_{13} - d_{23})$. We can use the variances of the estimated pair-wise

TABLE 1. Maximum Likelihood Estimates of Node Distances (*d*) and Node Ages (*t*) under Global and Local Clock Models

	Global clock		*Local clock*	
		Time		*Time*
1	$d_1 = 0.063$	$t_1 = 15.77 \pm 1.55$	$d_1 = 0.060$	$t_1 = 17.16 \pm 1.86$
2	$d_2 = 0.052$	$t_2 = 13$	$d_2 = 0.045$	$t_2 = 13$
3	$d_3 = 0.031$	$t_3 = 7.73 \pm 1.03$	$d_3 = 0.029$	$t_3 = 8.38 \pm 1.17$
4	$d_4 = 0.020$	$t_4 = 5.04 \pm 0.86$	$d_4 = 0.019$	$t_4 = 5.60 \pm 0.96$
5	$d_5 = 0.009$	$t_5 = 2.35 \pm 0.64$	$d_5 = 0.009$	$t_5 = 2.65 \pm 0.72$
l	-1796.12		-1795.35	
r_1	1		1.36	

Note. The data consist of the transfer RNA genes from five hominoid species: human, chimpanzee, gorilla, orangutan, and siamang. Alignment gaps are removed, with 759 nucleotides in the sequence. The model of nucleotide substitution accounts for different rates of transitions (T ↔ C or A ↔ G) and transversions (T, C ↔ A, G), as well as different frequencies for the four nucleotides. The local clock model assumes an independent rate r_1 for the orangutan lineage.

sequence distances to work out the standard error of the test statistic, σ_d, then compare d/σ_d with the standard normal distribution to test whether d is different from zero. This test is limited to three species only, and it does not test whether the outgroup species has a different rate from the two ingroup species.

A second test of the clock assumption is the likelihood ratio test (see Vignette). This test is applicable to data of any number of species. Figure 2 shows an example of testing the molecular clock using data of eleven transfer RNA genes in the mitochondrial genome from six hominoid species. The null model H_0 assumes the clock, and the parameters of the model are the $n - 1 = 5$ distances from the internal nodes of the tree to the present time, where n is the number of species in the tree (Figure 2B). The more complex model H_1 uses one independent rate parameter for each branch in the tree. Under this model, it is generally impossible to identify the root of the tree, so the "unrooted" tree is used, with $2n - 3$ branches (Figure 2A). Thus, model H_1 involves $(2n - 3) = 9$ parameters. The log-likelihood values under the clock and no-clock models are $l_0 = -1796.12$ and $l_0 = -1794.49$. We then compare $2\Delta_l = 2(l_1 - l_0) = 3.26$ with a χ^2 distribution with d.f. $= (2n - 3) - (n - 1) = n - 2 = 9 - 5 = 4$. The P value is 0.52, with d.f. = 4, and the clock is not rejected.

We should note that failure to reject the clock assumption does not necessarily mean that the evolutionary rate is constant over time. First, the null hypothesis tested by the likelihood ratio test is weaker than the assumption of a constant rate over time. For example, if the evolutionary rate has been accelerating over time in all lineages, the tree will look clocklike, although the rate is not constant. Furthermore, neither the likelihood ratio nor the relative rate tests can distinguish a variable from a constant rate within a lineage. Finally, failure to reject the clock might simply be because of a lack of information in the data rather than the correctness of the clock assumption. These observations suggest that the clock assumption should be accepted with caution when we estimate divergence dates from molecular data.

A third test is based on the index of dispersion, that is, the variance to mean ratio of the number of substitutions over lineages. When the rate is constant, the number of substitutions should follow a Poisson distribution, which has the mean equal to the variance. If all the species diverged at the same time and accumulate substitutions at the same rate since their divergence, the variance to mean ratio of the number of nucleotide substitutions among lineages should be close to one. Most real data sets generate dispersion indices much higher than one, an observation referred to as the overdispersed clock. Early studies by Kimura and J. Gillespie used gene sequences from different orders of mammals and assumed that they diverged in a radiation. This was later shown to be unacceptable as the phylogenetic relationship among species has much effect on the test. Recent analyses used only three lineages (primates, artiodactyls, and rodents) to avoid the problem of phylogeny, but the analysis is prone to errors as the variance calculation using only three lineages is unreliable. Tests based on the dispersion index are out of date and can be performed more rigorously using the likelihood ratio test.

Application of Molecular Clock to Estimate Divergence Dates. The most important practical application of the molecular clock hypothesis is to estimate divergence dates between species, populations, or even viral strains. Molecular sequence data allow estimation of distances only. Under the assumption of a constant rate over time, the distance is a linear function of time. To convert distances into absolute times, an external time (called a calibration point) is used, obtained from fossil data or geological events that mark the separation of species. In the example of Figure 2B, the clock assumption is used to estimate the divergence dates among the hominoid species, and the divergence time of the orangutan is fixed at thirteen million years ago for calibration. There are currently two major approaches to estimating the node distances, the distances from the internal nodes of the tree to the present time measured by the expected number of substitutions per site (the *d*'s). The first approach estimates the sequence distance between each pair of species, then uses least squares to fit the node distances to the pair-wise distance matrix. The second approach is maximum likelihood (Vignette), which is applied to the original sequence alignment rather than estimated pairwise distances. Estimates given in Table 1 were obtained using maximum likelihood. After the node distances are estimated, it is an easy matter to obtain the node times using the calibration point. The substitution rate can then be calculated as $\mu = 0.0517/(13 \times 10^6) = 3.976910^{-9}$ substitutions per site per year. We then use this rate to convert other distances into times, for example, $t_4 = 0.0201/(3.9769 \times 10^{-9}) = 5.04 \times 10^6$ years ago for the separation of humans and chimpanzees. Equivalently, $t_4 = 0.0201/0.0517 \times 13 = 5.04$ million years ago.

Note that both the calculation of pair-wise sequence distances and the likelihood joint analysis of all sequences rely on a model of nucleotide substitution, and it is important to use an adequate substitution model. A simplistic model does not correct for hidden changes (multiple substitutions at the same site) properly; as a result, the estimated distance will not be linear with time. For example, the proportion of differences between two sequences is not a linear function of time. It underestimates the distance, and the underestimation is more serious for large distances than for small ones, because more multiple substitutions are expected the longer the sequences have been diverged. This nonproportional underestimation of distances generates systematic biases in divergence date estimation.

Because the molecular clock hypothesis is often violated, especially for data sets of divergent sequences, much effort has been taken to estimate divergence dates even though the clock does not hold. Recent work has suggested it might be possible to estimate dates without assuming a global molecular clock. Two approaches have been taken. The first uses a random process to describe the change of the evolutionary rate over branches in the phylogeny and then use the Bayes method to estimate the posterior distribution of rates and dates. The second approach uses the likelihood method and assigns specific rate parameters to branches that are assumed to have different rates from other branches. For example, if we assume an independent rate for the orangutan lineage in Figure 2C, the maximum likelihood estimate of the human–chimpanzee divergence became 5.60 ± 0.96 million years ago, older than the estimate under the global clock model (Table 1). Although the clock assumption was not rejected by the likelihood ratio test, we note that date estimation is quite sensitive to assumptions about the clock.

BIBLIOGRAPHY

Cooper, A., and R. Fortey. "Evolutionary Explosions and the phylogenetic Fuse." *Trends in Ecology and Evolution* 13 (1998): 151–156. Discusses discrepancies between molecular dates and fossil records concerning two important periods in the evolutionary history: the Cambrian explosion and the origin and divergence of birds and mammals at the Cretaceous–Tertiary boundary.

Felsenstein, J. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *Journal of Molecular Evolution* 17 (1981): 368–376. Introduces maximum likelihood estimation of branch lengths and divergence dates as well as the likelihood ratio test of the clock assumption.

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge, 1983. Summarizes evidence and controversies concerning the molecular clock in relation to the neutral theory.

Thorne, J. L., H. Kishino, and I. S. Painter. "Estimating the Rate of Evolution of the Rate of Molecular Evolution." *Molecular Biology and Evolution* 15 (1998): 1647–1657. This paper Introduces the Bayes approach to date estimation using local molecular clocks.

Wu, C.-I., and W.-H. Li. "Evidence for Higher Rates of Nucleotide Substitution in Rodents Than in Man." *Proceedings of the National Academy of Sciences USA* 82 (1985): 1741–1745. Introduces the relative-rate test.

Yang, Z. "Among-Site Rate Variation and Its Impact on Phylogenetic Analyses." *Trends in Ecology and Evolution* 11 (1996): 367–372. Discusses date estimation using the molecular clock and the importance of evolutionary models.

Yoder, A. D., and Z. Yang. "Estimation of Primate Speciation Dates Using Local Molecular Clocks." *Molecular Biology and Evolution* 17 (2000): 1081–1090. Implements maximum likelihood models of local molecular clocks, and demonstrates the effects of substitution model, clock assumption, and calibration points on date estimation.

Zuckerkandl, E., and L. Pauling. "Evolutionary Divergence and Convergence in Proteins." In *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel, pp. 97–166. New York, 1965. Proposes the molecular clock hypothesis, among many important contributions.

— ZIHENG YANG

MOLECULAR EVOLUTION

Molecular evolution is a discipline of biology that utilizes molecular data to address evolutionary questions. The molecular data are usually DNA or protein sequences but may also include other types of data, such as the three-dimensional structure or some biochemical properties of a protein. DNA and proteins are referred to as *macromolecules* because they are much larger than molecules such as oxygen or ethanol. The field addresses diverse questions, ranging from traditional questions of life science to new questions driven by recently