“Molecular Clock” Analogs: a Relative Rates Exercise

John P. Wares
Department of Genetics, University of Georgia, Athens, Georgia 30602
Email: jpwares@uga.edu

Abstract: Although molecular clock theory is a commonly discussed facet of evolutionary biology, undergraduates are rarely presented with the underlying information of how this theory is examined relative to empirical data. Here a simple contextual exercise is presented that not only provides insight into molecular clocks, but is also a useful exercise for demonstrating how statistical processes are involved in modeling biological phenomena. The example given involves studying rate variation in traffic flow; a variety of other cases will be just as useful, and can provide founding material for further discussion of how molecular clock models are useful in basic and applied biology.

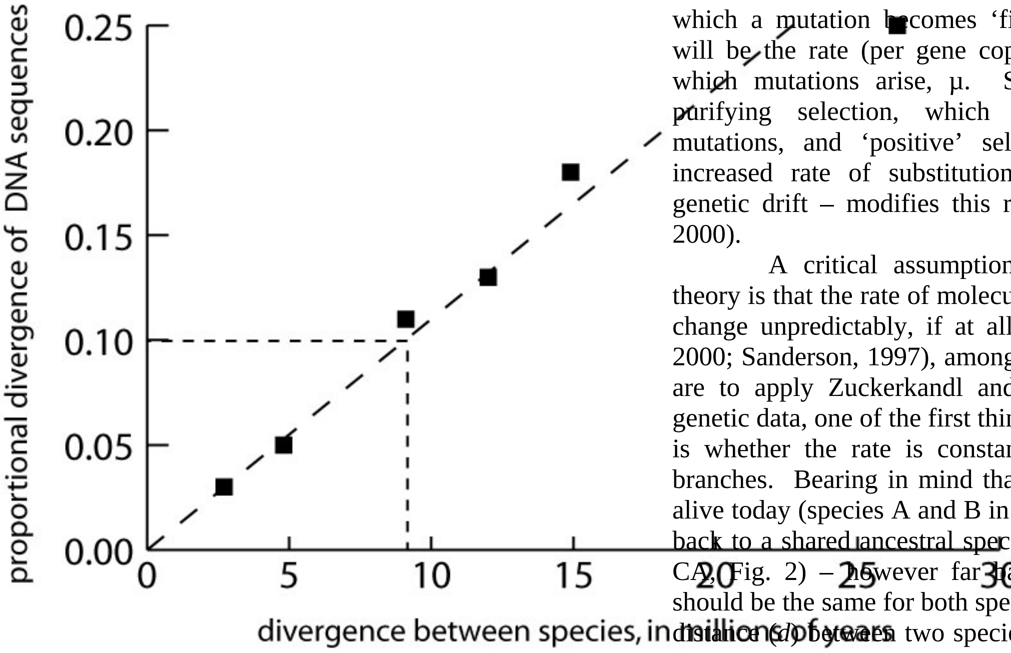
Keywords: molecular clock, relative rates test, Poisson distribution

Introduction

One of the more pervasive and often misunderstood aspects of evolutionary theory in popular science journalism is the “molecular clock”. This model, originally proposed by Zuckerkandl and Pauling (1965), revolved around a profound insight into how proteins differ among organisms: the more distantly related in time two organisms seem to be, the proteins they are composed of are also more distinct, suggesting that one process is an analog of the other. It is often discussed in science journalism, because it is a relevant and accessible feature of evolutionary biology: big discoveries often revolve around the age of a fossil, the time of separation of two lineages, and so on (Zimmer, 2003, 2005). This is also a point of contention in the debate between Creationists and evolutionary biologists. The former believe that clock estimates may represent circular logic (see Miller, 1999), requiring the inference of fossil dates to calibrate dates based on a molecular clock, while the latter are integrating information from diverse fields of science as well as testing the assumptions of clock models in a variety of ways (Pybus, 2006).

Thus, in undergraduate biology classes, it is often important to discuss the data and theory relevant to this model. A model is, after all, simply a testable way of describing what we see in nature. The broad implications of the molecular clock model are appealing – we can tell the age of an event by the molecular divergence of two lineages, assuming the clock has been ‘calibrated’ appropriately (Figure 1). Usually this involves information from breeding studies (Keightley and Lynch, 2003; O'Connell et al., 1997) or appropriate choice of calibration events – often from the fossil record or based on biogeographic events such as the rise of the Panama isthmus (Hickerson et al., 2003; Marko, 2002) – for a given taxon and era of interest. It can be difficult to establish robust divergence times using molecular clocks if the calibration points are an order of magnitude older or younger than the divergence of interest. Even after establishing appropriate rates for a gene region, and having data to address a particular question, it is important to remember that a certain amount of error is to be expected. The prominent evolutionary biologist Joe Felsenstein (Felsenstein, 2004, p.6-7) comments: “With a molecular clock, it is only the *expected* amounts of change [in two diverging lineages] that are equal; the *observed* amounts may not be.”

FIG. 1. An example of using a molecular clock model to estimate the divergence time of two species. This is useful when there is fossil, biogeographic, radiometric, or other means of dating some species divergences in a group, but not all of them. First, the “known” divergences between species are collected, and DNA sequence data is collected and compared for each of these species pairs (shown as black squares). A rate of mutation and substitution is inferred from a linear regression of these points. Then a comparison can be made between two species that have no information regarding their divergence time: sequence data is collected, and the divergence between those species compared against the regression line (shown as dotted line box). In this example, the ‘unknown’ species differ at 10% of the nucleotides in a DNA sequence; the inferred time of speciation is then about 9 million years ago. The variance in these estimates can also be accounted for statistically. (next page).



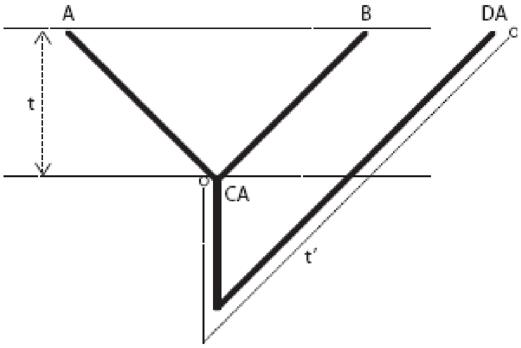
It is not straightforward to illustrate a mutational process in a teaching or laboratory setting, much less to show how we test mutational data to see whether they are appropriate for timing divergences among species. A challenge for lab classes in general, when dealing with evolutionary topics, is that evolutionary processes often require much more actual time than can be managed in the constraints of a classroom setting. Thus, analytical and simulation-based tools (e.g. EVOBEAKER by Simbiotic Software) are increasingly popular ways to allow students to see idealized processes happen. However, illustrating clock-like divergence is difficult in a computational lab, because it is in many ways an opaque process – a computer tells the students what is put into it. The abstraction of phylogeny means that simply showing unequal branch lengths (that is, the inferred number of mutations along a branch of a gene tree) may not help students learn both the statistical and model-testing concepts underlying molecular clock theory. Here I present an inexpensive and short lab exercise that can be easily modified to represent numerous real-world scenarios, to give students the opportunity to observe Poisson-distributed processes and determine whether an "equal rates" hypothesis is violated.

Theory

Kimura (1968) argued that most genetic variation – represented by different mutational alleles of a given gene in a population – must be neutral with respect to natural selection, in order to explain there being so much variation in natural populations. One of the consequences of this model is that the rate at which a mutation becomes 'fixed' in a population will be the rate (per gene copy, per generation) at which mutations arise, µ. Selection – including purifying selection, which removes deleterious mutations, and 'positive' selection that leads to increased rate of substitution relative to random genetic drift – modifies this rate (Hartl and Clark, 2000).

A critical assumption of molecular clock theory is that the rate of molecular evolution does not change unpredictably, if at all (Huelsenbeck et al., 2000; Sanderson, 1997), among species. Thus, if we are to apply Zuckerkandl and Pauling's theory to genetic data, one of the first things that must be tested is whether the rate is constant along evolutionary branches. Bearing in mind that for any two species alive today (species A and B in Figure 2), the time (*t*) back to a shared ancestral species (common ancestor CA, Fig. 2) – however far back in time that is – should be the same for both species. Thus the genetic distance (*d*) between two species is the mutation rate µ multiplied by the time *t* that has passed along *both* branches, or *d*=2µ*t*.

Figure 2. To examine whether the rate of evolution is constant in a group of species being compared, pairs of species (represented by DNA sequences, for example) are compared for divergence relative to a common ancestor (CA); by definition, the number of generations (time, *t*) separating each species from the common ancestor is the same. Often there will be no available sequence data from a common ancestor; a true outgroup sequence from an extant species, which has been diverged from the focal species (and their common ancestor) for a longer period of time, can be used to test the same relative rates hypothesis. If the substitution rate deviates dramatically between the outgroup (DA) and each species (A and B, below), other processes such as selection or demographic change may make estimates of divergence time based on a molecular clock model unreliable.



Without knowing μ, biologists can still evaluate the constancy of mutation using a relative-rates test. Since we often cannot sequence the DNA of an ancestral species (the common ancestor CA), we must compare the evolutionary divergence of each species to an 'outgroup' species. Although in general the molecular clock model performs very well for recent and ancient divergences alike, it is most common to only test the assumptions within a group that a researcher is interested in (say, turtles) and for a particular time scale (*e.g.* the phylogeny at the family level). Thus, DNA sequences can be separated into the 'ingroup' (the group that the researcher is considering) and the 'outgroup' species (those species that are known to be distinct from the ingroup – for example, a lizard could be an outgroup species for a phylogeny of turtles, as they are both in Reptilia but turtles share a more recent common ancestor with each other than with any lizard species; see Table 1).

TABLE 1. Researchers must often define outgroup species so that they understand the relationships within a group of species (the 'ingroup') better. These outgroup species must be from a distinct lineage of species, but preferably not so different that DNA substitution patterns are effectively random. Examples of outgroup species for several ingroups are given. The goal of choosing an outgroup for testing molecular clock theory is that for any two or more closely-related species, exactly the same amount of time must have passed since each of those species (which it is assumed – and can be tested – have a fairly recent common ancestor) had a more ancient common ancestor with a species that is from a distinct evolutionary lineage.

INGROUP	OUTGROUP
Genus *Quercus* (oak trees)	Maple tree, Poplar tree
Order Cirripedia (barnacles)	Crab, Shrimp
Genus *Oncorhynchus* (trout)	Atlantic Salmon (genus *Salmo*)
Family Ursidae (bears)	Dogs
Phylum Echinodermata (seastars)	Sea squirts

The basis of the relative rates test (Sarich and Wilson, 1973), and all of the more complex variants that are used in examining the time of species divergence based on genetic distance (Sanderson, 1997; Huelsenbeck et al., 2000; Pybus, 2006), is the underlying statistical distribution. The rate expectation is based on a very simple statistical distribution that is worth discussion in any science class or laboratory – the Poisson.

Mutations occur randomly over time and are often modeled by a Poisson distribution. The Poisson distribution is usually employed for modeling systems where the probability of an event occurring is very low, but the number of opportunities for such occurrence is very high. Generally, the probability of mutation in a single generation is low, but over evolutionary time there are many thousands of generations separating two species or even individuals within a species. The basic assumptions of a Poisson distribution are:

(1) the length of observation period is fixed in advance
(2) events occur at a constant average rate
(3) the number of events occurring in different intervals is independent.

Thus typical examples that are given include the waiting time or frequency with which light bulbs burn out, because we know it will happen but the waiting time could be quite short or quite long, keeping usage and other conditions constant. A difficulty with making such a concept, and its statistical underpinnings, understandable to students is finding an example that is observable, quantifiable, and useful in the context of observing differential rates, as with the mutational changes on two distinct evolutionary lineages.

Exercise

This exercise should begin with a short lecture explaining the underlying theory (previous section) and examination of typical applications of molecular clock data (such as Zimmer, 2003, 2005), including information from the fossil record and divergence data from DNA sequences. There is ample material available online and in the literature for this portion of the exercise (*e.g.* Kalinowski et al., 2006). The participatory focus of the exercise will be adapted to local surroundings, but the suggested opportunity involves finding a nearby area of high vehicle or pedestrian traffic in which people (vehicles, organisms, objects) are forced to turn either to the right or the left, as with a T-intersection at a traffic light (Figure 3). The events should not be immediately predictable as to whether the turn will be left or right, and one reason that choosing an intersection with a stoplight is useful is because the light breaks the number of events into (semi-) discrete intervals, appropriate for examining a Poisson distributed process.

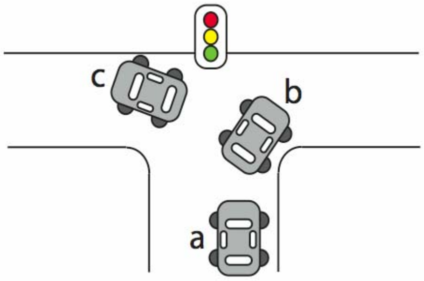
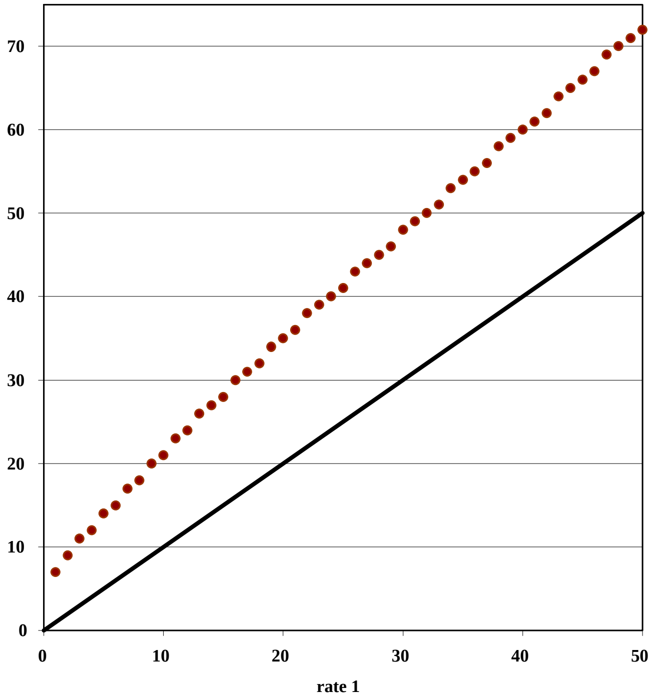


FIG. 3. Traffic scenario for illustrating Poisson processes and an analog to the molecular clock model. Here, the car at **a** is in the ancestral (prior) state; car **b** has turned to the right, indicating an independent mutation that distinguishes it from the ancestral state and from cars that turn to the left (**c**).

In the case of a T-intersection, students are given a portion of the lab time to safely observe and record the rate of turns in either direction. They are invited to subdivide the data into different partitions, such as 'cars' versus 'trucks' or 'university vehicles', which may have distinct rates from one another – much as different loci may evolve in different ways in the genome. After a series of time intervals appropriate for the rate of traffic flow (see Figure 4), in which all left and right turns for each partition (*i.e.*, vehicle type) are recorded in their lab notebooks, students return to the lab or a setting in which they can calculate the rate variation on each 'branch' being observed.

The students will use what is called a relative rates test to determine whether the rate is the same for both 'branches'. Ordinarily, this would be done in the context of having genetic data for at least one additional species (the outgroup). However, with these data the test statistic can be calculated in exactly the same way, as though turns to either direction represent a branch from the ancestral position to each ingroup destination (species). Tajima (1993) showed that if there are three nucleotide sequences then we can define the number of sites in which nucleotides in sequence 1 (one of the two ingroup species) are different from the other two sequences by m_1. The number of sites in sequence 2 that are different from the other sequences is defined as m_2. When sequence 3 is considered the outgroup, the expectation for equal rates is that $m_1 = m_2$.

FIG 4. The χ^2 statistic used in relative rates tests becomes more sensitive as sample size increases. For very small numbers of observations, the excess of events for one of the rates being measured must be much larger than the other rate; as the sample size (number of observations) increases, the ratio between the rates can be smaller and still be statistically significant. Shown is a diagonal line illustrating equal rates, and dots indicate for a given number of observations for one rate estimate the number of discrete events in the other rate estimate that will be significant ($\chi^2 = 3.84$) assuming 1 degree of freedom ($p < 0.05$).



This equality can be tested using the chi-squared distribution. Namely,

$$\chi^2 = \frac{(m_1 - m_2)^2}{m_1 + m_2}$$

approximately follows the chi-square distribution with one degree of freedom. This test is conservative, meaning that it will not always have the power to detect rate variation, but it is a good first approach. The students will use their data, where left turns are equivalent to unique substitutions in sequence 1, and right turns are equivalent to unique substitutions in sequence 2, to determine if there is statistically significant rate variation (here, if the test statistic is > 3.84 it would be considered significant at the $p = 0.05$ level; this also offers an opportunity to discuss what is meant by 'significance' in scientific tests).

Discussion

An important element of any teaching exercise is finding a way to make the subject matter memorable. It is very difficult to establish ways to educate students about statistical distributions, and particularly as related to rate variation, in a way that involves activity. While there is much to be gained from discussions and simulation-based exercises, abstract concepts are better conveyed through visual and participatory activities (Kalinowski et al., 2006). This exercise has been reviewed as one of the more memorable labs in my own teaching, and students did well on subsequent exam questions related to rate variation and the molecular clock. Students learn that some data partitions (*e.g.* University vehicles) may exhibit significant rate variation relative to other vehicle types, and that when there is significant rate variation the data in question may not be useful for examining questions that involve an *assumption* of constant rate, such as is used in molecular clock models. They also learn how scientists examine their own data in evaluating whether a particular model may be applied for further evaluation.

There are abundant instances in which researchers have found that the data they have collected are not consistent with a constant clock-like rate of evolution. In many cases, using relative-rates tests such as the one described and illustrated above is a way of examining why the data are inconsistent; for example, Posada (2001) showed that recombination within gene regions can often lead to larger variances in branch lengths among species and rejection of a molecular clock model. In these instances, other means of inference are required to determine the time of separation for species, or the rate at which other characters are changing in the course of evolution. However there are a great many cases in which DNA sequence data is able to accurately predict the divergence time among lineages, even on very short time scales if the rate of mutation is quite high. Nickle et al. (2002) found that the rate of sequence evolution in HIV samples fits a molecular clock model; this finding is clinically relevant as it may help infer sources of infection and transmission in a patient's history. More work is being done to understand why some data sets deviate from the expectations of the model, and how to consider factors such as whether the rate can change over time (Huelsenbeck, 2000).

Unusual activities such as the one outlined here provide a more memorable experience when teaching abstract topics. As noted in Kalinowski et al. (2006), participatory exercises are often more effective and memorable than complementary approaches to teaching the same subject matter, particularly when the topic is relatively abstract in nature. Although the process described in this exercise is not a perfect analog to the process of mutation in independent lineages – for example, it is not possible for a car to turn left as well as right, but it is possible for a single nucleotide to mutate in two independent species lineages (a condition known as *homoplasy* or *parallel mutation*) – many alternative scenarios may be explored for such an exercise. This is presented primarily to illustrate ways in which educators can take advantage of local conditions to teach basic probability theory, to help explain a topic related to evolution and the molecular clock model, and to avoid the 'black box' problem of some computer-based simulation exercises.

Acknowledgments

Thanks to Elizabeth Jockusch for sharing the molecular clock exercise that has been used in my own teaching lab along with the above exercise, and Scott Small for assisting with the Evolutionary Genetics teaching lab at the University of Georgia.

References

FELSENSTEIN, J. 2004, Inferring Phylogenies, Sinauer.

HARTL, D. L. 2000, A primer of population genetics. Sunderland, MA, Sinauer Associates.

HICKERSON, M. J., M. GILCHRIST, AND N. TAKEBAYASHI. 2003. Calibrating a molecular clock from phylogeographic data: moments and likelihood estimators. *Evolution* 57:2216-2225.

HUELSENBECK, J. P., B. LARGET, AND D. SWOFFORD. 2000. A compound process for relaxing the molecular clock. *Genetics* 154:1879-1892.

KALINOWSKI, S. T., M. L. TAPER, AND A. M. METZ. 2006. How are humans related to other primates? A guided inquiry laboratory for undergraduate students. *Genetics* 172:1379-1383.

KEIGHTLEY, P. D., AND M. LYNCH. 2003. Toward a realistic model of mutations affecting fitness. *Evolution* 57:683-685.

KIMURA, M. 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res*. 11:247-269.

MARKO, P. B. 2002. Fossil calibration of molecular clocks and the divergence times of geminate species pairs separated by the Isthmus of Panama. *Mol. Biol. Evol.* 19:2005-2021.

MILLER, K. 1999, *Finding Darwin's God*. New York, HarperCollins.

NICKLE, D. C., Y. LIU, G. H. LEARN, D. SHINER, AND J. MULLINS. 2002, HIV Evolution is Largely Consistent with a Molecular Clock. 9th *Conference on Retroviruses and Opportunistic Infections*.

O'CONNELL, M., R. G. DANZMANN, J. CORNUET, J. M. WRIGHT, AND M. M. FERGUSON. 1997. Differentiation of rainbow trout (*Oncorhynchus mykiss)* populations in Lake Ontario and the evaluation of the stepwise mutation and infinite allele mutation models using microsatellite variability. *Can. J. Fish. Aquat. Sci.* 54:1391-1399.

POSADA, D. 2001. Unveiling the molecular clock in the presence of recombination. *Molecular Biology and Evolution* 18:1976-1978.

PYBUS, O. G. 2006. Model selection and the molecular clock. *PLoS Biology* 4:e151.

SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol BIol Evol* 14:1218-1231.

SARICH, V. M., AND A. C. WILSON. 1973. Generation time and genomic evolution in primates. *Science* 179:1144-1147.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607.

ZIMMER, C. 2003. Adam's Family, *The New York Times.* New York City.
—. 2005. Plain, simple, primitive? Not the jellyfish, *The New York Times*.

ZUCKERKANDL, E., AND L. PAULING. 1965. Molecules as documents of evolutionary history. *J. Theoret. Biol.* 8:357-366.