Estimating Rate and Time in Molecular Phylogenies: Beyond the Molecular Clock?

Michael J. Sanderson

The study of rates of character evolution has been a cornerstone of evolutionary biology since the pioneering work of Simpson (1944). It has occupied a similar position in molecular evolutionary studies since Zuckerkandl and Pauling’s (1962, 1965) proposal of the molecular clock. There is a fascinating contrast between these two works, however. Simpson used information about time, from the fossil record, to draw inferences about rates and modes of evolution. His main conclusion was that such rates are highly variable. Although also using information from fossils, Zuckerkandl and Pauling came to just the opposite conclusion about rates of protein evolution. They then argued that if proteins evolved at a roughly constant rate, a study of rates and modes of evolution could be used to say something about timing of events in evolutionary history. Both these ideas about the tempo of character evolution have achieved nearly the status of null hypotheses in their respective disciplines. Although Simpson clearly inferred that some morphological rates have been nearly linear, or “clock-like” over at least moderate periods of time (e.g., Simpson, 1944, pp. 203–204), few paleontologists or morphologists give credence to the notion of morphological clocks. And although there is indisputable evidence that many genes and proteins do not evolve at a constant rate through time (Britten, 1986; Avise, 1994), molecular rate constancy continues to be viewed as a reasonable model even across vast reaches of the tree of life (Wray et al., 1996).

One obvious reason for this difference is that there was no neutral theory that predicted rate constancy at the morphological level, and thus never any process-based justification for using morphological divergence to date events in history. The pervasiveness of the null model of molecular rate constancy, however, led to many reconstructions of timings of key evolutionary events, often without even a passing attempt to test for the existence of a clock, even in cases where the neutral theory might not be expected to hold, such as in nonsynonymous substitution rates. A classic example isRamshaw et al.’s (1972) reconstruction of an angiosperm age of 300–400 Ma based on amino acid replacement rates in cytochrome c, clearly at odds with the Early Cretaceous origin strongly suggested by the fossil record (Doyle and Donoghue, 1993).

Despite this contrast there are parallels between the development of the study of morphological and molecular rates of evolution. Just as the fossil record led Simpson to conclude that rates could vary over time and across taxa, evidence left in genes and proteins eventually led molecular evolutionists to conclude the same.

I thank Steve Nadler, the editors and their students, and an anonymous reviewer for comments on the manuscript.
thing about rates at the molecular level. It was also clear from the outset that rates varied among regions of genes and proteins subject to constraints imposed by the genetic code, protein structure or function, expression, and so on. This evidence continues to play a key role in testing theories of evolution at the molecular level, including neutralist versus selectionist models (Kimura, 1983; Gillespie, 1991). Tests of rate constancy across lineages became possible even in the absence of a fossil record using relative rate tests (Sarich and Wilson, 1967), and considerable heterogeneity across lineages was discovered (Britten, 1986; Avise, 1994). Notable cases included the higher rate in rodents versus primates (Wu and Li, 1985), grasses versus palms (Wilson et al., 1990; Gaut et al., 1992), and annual versus perennial angiosperms (Savard et al., 1994).

The literature on rates has been reviewed with respect to the neutral theory (Gillespie, 1986, 1991) and with respect to phylogenetic inference and the many potential biases that phylogenetic practice must contend with (Mindell and Thacker, 1996). Reviews of the empirical literature have been infrequent and not comprehensive (Britten, 1986; Wolfe et al., 1987; Gaut et al., 1993; Avise, 1994). Estimation of divergence times in particular has been reviewed briefly by Springer (1995), and Hillis et al. (1996), who both emphasize calibration issues. However, there has not been a comprehensive review of the methodology for assessing rate variation across lineages and for estimating divergence times. Many of these methods have been developed within the last 5 years. This paper reviews methods aimed at estimating lineage specific rates of evolution and/or divergence times. It does not consider the equally rich and biologically interesting set of issues associated with estimating rates across different entities within the same lineage: that is, rates in different genes, different sites, synonymous versus non-synonymous rates, and so forth. Some of these issues can be easily built into the models of molecular evolution used to study lineage variation in rate (e.g., third position versus first or second positions in codons), but others raise technical problems that are daunting in their own right, such as a truly correct inclusion of synonymous versus nonsynonymous rate parameters in models of protein coding sequences (Li et al., 1985; Goldman and Yang, 1994; Muse and Gaut, 1994).

Some methods of estimating evolutionary rates and divergence times require a phylogeny; others do not. If a phylogeny is needed for a particular method, it will be assumed that an algorithm is available to provide it (e.g., Swofford et al., 1996). The error introduced into the estimation of rates or divergence times because of inaccurate phylogenetic estimation is not considered further.

DEFINITIONS OF TERMS

A rate is a change in some quantity in an interval of time. An absolute rate specifies the time scale, such as numbers of substitutions per million years, whereas a relative rate merely compares two quantities observed during the same time interval, such as the inferred number of substitutions in two sister lineages known to be the same age. In general, a rate can be thought of as a property of some deterministic or random process that proceeds in time. Thus it might be an average taken over many discrete time intervals, or it might be a parameter of some very detailed model of evolution that prescribes the instantaneous rate of change of DNA sequences. In either case it is usually something that is estimated from data via a statistical procedure for extracting signal from noise.

The various meanings of branch length have caused endless confusion. Branch length refers to the amount of character-state change occurring along a branch. This may be an integer in the case of discrete characters, or a real number in the case of genetic distances that are continuous valued. In trees reconstructed with the aid of a probabilistic model, branch lengths usually refer to the expected (i.e., mean) number of changes occurring as an outcome of some probability of change over some duration. Less commonly in the empirical literature but not infrequently in theoretical papers (Cavender, 1978), branch length refers simply to the probability of states differing at the endpoints of a branch. As such it is a real number between 0 and 1. This meaning will not be used further. Length should