Review

Watching the clock: Studying variation in rates of molecular evolution between species

Robert Lanfear[1], John J. Welch[2] and Lindell Bromham[1]

[1] Centre for Macroevolution and Macroecology, Evolution Ecology and Genetics, Research School of Biology, The Australian National University, Canberra ACT 0200, Australia
[2] Université Montpellier 2, Place Eugène Bataillon, Montpellier, France

Evidence is accumulating that rates of molecular evolution vary substantially between species, and that this rate variation is partly determined by species characteristics. A better understanding of how and why rates of molecular evolution vary provides a window on evolutionary processes, and might facilitate improvements in DNA sequence analysis. Measuring rates of molecular evolution and identifying the correlates of rate variation present a unique set of challenges. We describe and compare recent methodological advances that have been proposed to deal with these challenges. We provide a guide to the theoretical basis and practical application of the methods, outline the types of data on which they can be used, and indicate the types of questions they can be used to ask.

Why study the rate of molecular evolution?

DNA sequences evolve at different rates in different species. Indeed, contrary to hopes that molecular evolution would be clock-like, variation in evolutionary rates between species appears to be the rule, rather than the exception [1–3]. A significant component of this variation is associated with species biology (see Box 1, [1,3–5]). For example, many studies have found evidence that species with shorter generation times tend to have faster rates of molecular evolution (e.g. [3,6–8]). Studying the factors that influence the rate of molecular evolution is important for understanding some of the most fundamental aspects of evolutionary biology, such as the relationship between genomic change and speciation [9,10], the link between molecular and morphological evolution [11,12], and the effects of species' life-history on evolution [1,3–6,13]. Furthermore, illumination of the causes of molecular evolution could play an important practical role in improving DNA sequence analysis. Molecular data are increasingly used in all areas of biology, and many analyses make assumptions about molecular rates. For instance, all molecular dating methods rely on assumptions about how molecular rates change over evolutionary time (reviewed in Ref. [14]). Our understanding of molecular evolution is often insufficient to know whether these assumptions are reasonable [15], and in some cases it has been clearly shown they are not [16]. To improve these methods, we will need to know how and why rates of molecular evolution vary between species.

Once upon a time, empirical studies of rate variation were limited by a lack of molecular data since DNA sequences were available for only a handful of species (e.g. [7,13,17]). Now, the availability of DNA sequences for hundreds of thousands of species and a growing number of databases of species life-history characteristics (e.g. [18]) has seen an increase in the number of studies testing links between species' traits and rates of molecular evolution.

Here, we review advances in methods for studying the association between the rate of molecular evolution and other biological factors, and discuss how these methods can

Glossary

Alignment: a representation of two or more homologous DNA (or protein) sequences in which each row represents a different sequence, and each column represents a single homologous nucleotide (or amino acid).
Branch length: an estimate of the number of substitutions that have occurred along a given branch of a phylogenetic tree, usually measured in substitutions per site of the alignment.
Nodes: branching points on a phylogenetic tree, which represent the last common ancestor of two or more lineages.
Non-synonymous substitution: a substitution in a protein-coding DNA sequence that changes the encoded amino acid.
Polymorphism: a change in the DNA sequence that is present in some, but not all, individuals of a species.
Rate-smoothing: a molecular dating method in which rates can vary among lineages in an auto-correlated fashion, i.e. incorporating an *a priori* assumption that closely related lineages are likely to have similar rates of evolution.
RY-coding: replacing the nucleotide letters of a DNA sequence so that purines (adenines and guanines, or As and Gs) are represented as Rs, and pyrimidines (cytosines and thymines, or Cs and Ts) are represented as Ys.
Substitution: a change in the DNA sequence that is present in virtually all individuals of a species.
Substitution rate: a measure of the number of substitutions that have occurred per unit time, usually measured in substitutions per site of the alignment per million years.
Synonymous substitution: a substitution in a protein-coding DNA sequence that does not change the encoded amino acid.
Terminal branch: a branch on a phylogeny that connects an observed DNA sequence to an internal node.
Terminal node: a node on a phylogeny from which no further nodes arise (otherwise known as a leaf node).
Transitions: a change in a DNA sequence from a pyrimidine to a pyrimidine (e.g. C to T), or from a purine to a purine (e.g. G to A).
Transversion: a change in a DNA sequence from a pyrimidine to a purine (e.g. G to T), or from a purine to a pyrimidine (e.g. A to C).
Uncorrelated relaxed molecular clocks: molecular dating methods in which rates can vary among lineages, but which do not assume that closely related lineages are more likely to have similar rates (as distinct from rate smoothing methods).

Corresponding author: Lanfear, R. (rob.lanfear@anu.edu.au).

0169-5347/$ – see front matter doi:10.1016/j.tree.2010.06.007 Trends in Ecology and Evolution 25 (2010) 495–503

Box 1. Using comparative methods to understand mammalian molecular evolution

Early comparisons of amino acid sequences or DNA hybridisation distances between species revealed that rodents had faster rates of molecular change than artiodactyls, which had faster rates than primates (e.g. [56,57]). The accumulation of sequence data from many species has allowed the generality of patterns of species differences in rates to be explored, revealing a trend in rates of molecular evolution with body size in vertebrates (e.g. [13]). The search for the cause of this body size trend has been challenging, for several reasons. There are many aspects of mammalian biology that could influence molecular rates, and most have phylogenetic inertia: the more closely related species are, the more similar they tend to be. Furthermore, in mammals, many of these traits scale with body size. For example, the observation that smaller mammal species tend to have faster molecular rates has been interpreted as evidence for the influence of metabolic rate on rates of molecular evolution (e.g. [58]), but smaller-bodied species also tend to have larger populations, faster turnover of generations, higher fecundity and shorter lifespans. If any of these life history variables affect molecular rates, there would be a spurious relationship with all of the others, making it a challenge to untangle the causes of rate variation.

Both phylogenetic inertia and covariation of different life history traits have been examined using phylogenetic comparative methods. An early comparative study of DNA sequences from phylogenetically-independent sister pairs of mammals showed that rates did scale with body size and generation time, but metabolic rate had no explanatory power beyond its association with the other life history traits [59]. Further studies, using more data, better estimates of rates, and more sophisticated comparative methods have shown that body size, generation time and fecundity scale with substitution rates in the nuclear genome, but that the body size trend in mitochondrial substitution rates appears to be wholly explained by variation in longevity, with longer-lived species having lower substitution rates [1,5]. One possible interpretation of these results is that species that copy their germline DNA more often per unit time (those with fast generation turnover and high reproductive outputs) accumulate more DNA copy errors in their nuclear genomes. Additionally, longer-lived species might have stronger selective pressure to reduce the lifetime mutation rate, particularly in the mitochondria, where DNA damage has been associated with age-related decline (reviewed in Refs [60–62]).

be used to improve our understanding of molecular evolution. To study rate variation, we need to be able to do two things. Firstly, we need to accurately estimate rates of molecular evolution. Secondly, we need a statistical framework for testing hypotheses about the causes of lineage-specific rates. Both of these things are trickier than they first appear.

Estimating substitution rates

Background

In most cases, the evolution of DNA sequences cannot be witnessed directly, so rates of molecular evolution must be estimated by comparing sequences from different species. To estimate a rate of molecular evolution we need to know how many substitutions have occurred in a given lineage during a known time period. To do this we must estimate the number of substitutions that have occurred in different lineages (branch lengths) using a model of molecular evolution.

The most common method of calculating branch lengths is to estimate them simultaneously with the topology of the phylogenetic tree. However, branch lengths can also be estimated on an 'assumed phylogeny' in which relationships between species are fixed based on some prior knowledge. Branch lengths are most commonly estimated using Maximum Likelihood (e.g. [19–21]) or Bayesian methods (e.g. [22,23]) (importantly, parsimony is not an appropriate method for estimating branch lengths from DNA or protein sequences because it does not allow for any site to have had more than one substitution on a single branch of the phylogeny).

Some methods for testing the correlates of rates require only branch lengths in order to estimate differences in substitution rates (e.g. sister pairs, see below). However, if the estimate of phylogeny also includes dates of divergence (see Figure 1 and below), then it is possible to estimate absolute substitution rates, typically in units of substitutions per site per million years. Absolute substitution rates can be compared directly among a wide range of taxa, so these rate estimates can be used with a wider range of methods for detecting the correlates of substitution rates (see below).

Getting the best estimates of substitution rates

Like any statistical estimate, the accuracy of substitution rate estimates depends on the adequacy of the model used, and the quality and quantity of the data (methods for choosing an appropriate model are reviewed in Refs [24,25]). It should not be forgotten that the most critical stage of any analysis of substitution rates is sequence alignment, as this provides the raw material for rate estimation. In addition to alignment, substitution rate estimates have some special properties that require attention, particularly issues associated with having too few or too many substitutions to make reliable estimates. The amount of sequence data needed to make reliable estimates will therefore depend on the rate of molecular evolution, the age of the divergences between species, and the way that substitutions have been distributed across sites in the sequence.

Estimating substitution rates can be tricky when comparing sequences with a small number of observable differences, which can be the result of slow rates, recent divergences or short sequences. When the total number of substitutions is small, the error in substitution rate estimates can be very high, because a small difference in the number of substitutions will represent a large proportional difference in the estimated substitution rate. In a comparative framework in which the change in substitution rate is the variable of interest (see below), this effect can be particularly problematic. A test has been proposed that uses patterns of variance in the estimated changes in substitution rates to identify misleading estimates of rate change [26]. Once identified, these imprecise estimates should be excluded from analyses of substitution rate variation. Although it might seem counterintuitive to remove data points from an analysis, it has been shown both in theory and in practice that this approach can greatly improve the power of subsequent statistical analyses to detect the correlates of substitution rate variation [5,26].

A second issue that can arise when the number of substitutions is small involves estimating rates for closely related species. Not all differences between a pair of

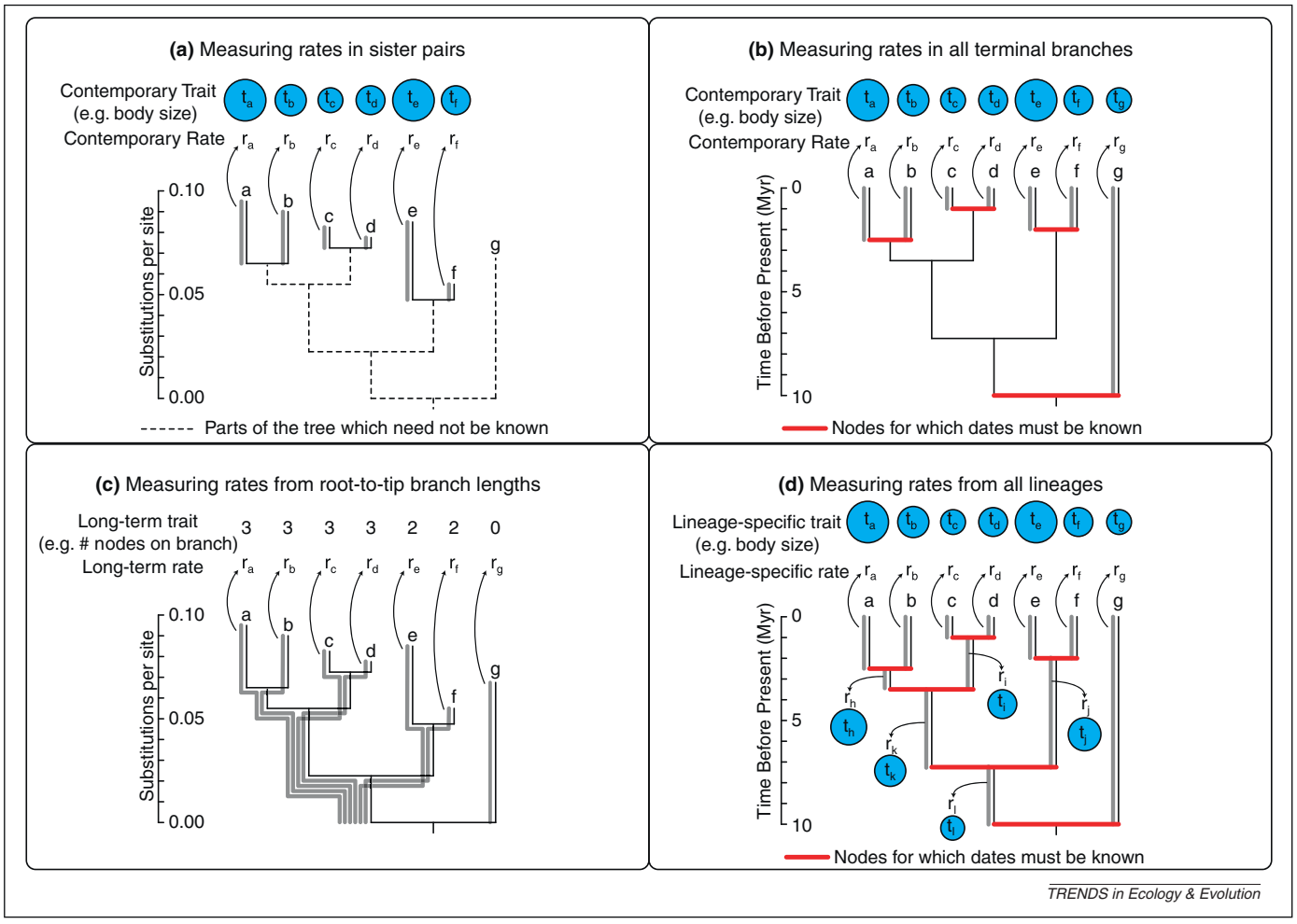

Figure 1. Four approaches to rate measurement in the comparative analyses of variation in substitution rates. The aim of all four methods is to generate estimates of substitution rates which can be used to test for associations between rates and traits using phylogenetic comparative methods. The underlying phylogeny is identical in all examples, but the methods differ in the way that substitution rates are estimated. The grey lines on the phylogenetic trees indicate those parts of a lineage's evolutionary history that are used to estimate evolutionary rates. The sister-pairs **(a)** and terminal-branch **(b)** methods estimate rates from the terminal branches of the phylogeny. These are treated as contemporary rate estimates, and compared to contemporary traits (such as body size measured from extant populations). The sister-pairs method **(a)** does not require dated nodes, and much of the phylogeny need not be known with confidence (shown in dotted black lines). The terminal-branch method **(b)** requires that the entire phylogeny and the relative dates of terminal nodes (shown in red) are known. The root-to-tip **(c)** and all-lineages **(d)** methods estimate rates from the entire phylogeny. The root-to-tip method **(c)** measures rates over the entire history of a lineage, and is suitable for comparing rates to similarly long-term traits, such as the number of nodes through which a lineage has passed. The all-lineages method **(d)** uses rate and trait estimates from every branch of the phylogeny, and is most appropriate when the dates of all nodes in the phylogeny are known, along with ancestral values of the trait of interest.

sequences will be substitutions derived from mutations which have occurred since reproductive isolation; some might be polymorphisms that happen to be present in the sampled individuals, or ancestral polymorphisms that have fixed in one or other lineage [27]. Mistaking polymorphisms for substitutions is a problem because the two quantities can vary with species characteristics in quite different ways. For example, the number of polymorphisms varies with population size, while the rate of neutral substitution does not [28]. A simple solution to this problem is to restrict analyses to divergences that are old enough for the number of polymorphisms to be negligible compared to the number of substitutions between species.

Getting accurate rate estimates can also be difficult when so many substitutions have occurred that sequences become saturated with changes, such that new substitutions overwrite old ones and erase their historical signal [29]. There are two common ways of mitigating the problem of saturation. First, data points based on rate estimates from saturated data can be excluded from the analysis (e.g. [30]). Second, substitution rates can be estimated exclusively from rare types of substitution, which tend to become saturated more slowly. For instance, substitution rates can be estimated only from non-synonymous substitutions (by using alignments of amino acids for example), which tend to occur more rarely than synonymous substitutions (e.g. [4]). Some researchers have also estimated substitution rates only from transversions (which tend to be rare), and not transitions (which tend to be more common), using a process known as RY-coding [4,31]. Where there is saturation in the dataset, it may be preferable to re-estimate all branch lengths from rare substitutions rather than remove the problematic estimates from the analysis. However, because different types of substitution scale with biological variables in different ways (see Box 2), using only rare types of substitution might not be appropriate in all cases.

For many branch-length estimation methods, including Maximum Likelihood, there can be an artefactual positive

Box 2. Dissecting the causes of substitution rate variation using DNA sequences

Different biological processes can affect rates and patterns of substitutions in different ways. These differences can be exploited to disentangle some of the hypotheses about variation in rates of molecular evolution.

In protein-coding sequences, synonymous mutations tend to have little or no effect on the fitness of the carrier [63], whereas non-synonymous mutations can have a much larger range of effects, from being lethal to strongly advantageous [64]. As a result of these differences, many processes are expected to affect the synonymous substitution rate (*dS*) to a different extent than they affect the non-synonymous substitution rate (*dN*). For instance, a change in the mutation rate will affect the appearance of synonymous and non-synonymous mutations equally. However, while the fixation of synonymous mutations will often be unaffected by natural selection (but see Ref. [63]), the fixation of non-synonymous mutations will depend on natural selection to a much larger extent. Because of this, a change in the mutation rate is more likely to leave a reliable signal in *dS* than in *dN* [1,4,5,10]. Additionally, changes in effective population size (N_e) affect the balance of power between genetic drift and natural selection, and so will often have more severe effects on *dN* than *dS* (reviewed in Ref. [65]). Because of this, the ratio of *dN* to *dS* (ω) has often been used to test whether variation in substitution rates can be explained by differences in N_e (e.g. [66,67]).

dN, *dS* and ω can be calculated from alignments of protein coding DNA using freely-available software such as HyPhy [68] and PAML [69]. However, differences in the way that these quantities are estimated can lead to important differences in the resulting values [70–72]. In particular, it has been suggested that values of *dS* and ω will usually be most accurately estimated by models which explicitly account for variations in *dS* across both sites and lineages [73], such as the 'Dual Rate-Variation' models implemented in HyPhy [68].

correlation between the number of nodes through which a lineage passes, and the estimate of the number of substitutions along that lineage. This is known as the node-density effect (e.g. [32]). This effect occurs because branch-length estimation algorithms tend to underestimate the number of substitutions that have occurred on long branches. The node-density effect can lead to artefactual patterns in rate variation, in which lineages that go through more nodes on the tree appear to have higher rates of molecular evolution. This has led to controversy around some studies of substitution rates, particularly those that examine links between diversification and rates of molecular evolution [33–36]. A test has been proposed which allows investigators to assess whether a given molecular phylogeny suffers from the node-density effect [33,37], and it has been used to exclude some molecular phylogenies from certain analyses [9]. However, the use of this test to exclude molecular phylogenies in this way remains controversial [32,38]. Because of this, it is preferable where possible to ensure that all substitution rate estimates are calculated from terminal branches of the tree (see e.g. Figure 1a and b; and methods 1 and 2, below), in which case the node-density effect cannot bias substitution rate estimates.

Finally, many methods for detecting the correlates of substitution rates require the calculation of absolute substitution rates, so they require that the relative dates of divergence of nodes in the tree are known. Ideally, these divergence date estimates should use sources of information other than the molecular data, although in practice, these dates are often calculated from molecular data themselves (e.g. [1]). If molecular dates are used to estimate rates, it is important to remember that all methods for estimating rates and dates from molecular data make assumptions about the way substitution rates can change over evolutionary time (reviewed in Ref. [14]). For instance, many molecular dating methods assume that substitution rates are likely to be similar in closely related lineages (e.g. rate-smoothing [39]). Although a number of studies of the correlates of substitution rates have employed such methods (e.g. [1,3,40]), it has been argued that this assumption might not always be justified, and that in many cases it might be preferable to use methods that do not make this assumption *a priori* (such as uncorrelated relaxed molecular clocks) [41].

Explaining variation in substitution rates

Background

When estimating rates of molecular evolution, the estimates are usually not identical for all lineages. Some of this variation might be due to the inaccuracy of evolutionary rate estimates (e.g. [42]), but some of the variation might be linked to differences in species biology. In this section, we describe methods that can be used to test for links between rates of molecular evolution and other aspects of species biology.

One of the most important aspects in designing analyses of rate variation is dealing with shared evolutionary history using phylogenetic comparative methods. Most statistical methods assume that data-points are independent of one another, but the fact that more closely-related species will tend to have more similar characteristics introduces non-independence into most datasets (see Box 3). Phylogenetic comparative methods are well established [43–48], and will usually be appropriate for the analysis of variation in substitution rates (but see [49]). The choice of which comparative method to use depends upon how much data is available, what is known about the underlying phylogenetic relationships, and how rates of molecular evolution have been measured (see Table 1). Below, we divide these methods into classes according to the parts of the underlying phylogeny that are used to estimate substitution rates.

Methods that use rate estimates from sister pairs The simplest approach that can be used to control for phylogenetic non-independence is the calculation of differences in rates and traits between pairs of sister taxa (Figure 1a). Each sister pair comprises two clades that share a common ancestor to the exclusion of all other such pairs in the tree (Figure 1a). Differences between the rates and traits of the two members of each pair are typically treated as statistically independent observations.

The sister pairs method is attractive because it requires very few assumptions to be made about the data. For instance, it is not necessary to have a fully resolved phylogeny as long there is confidence that the lineages connecting the members of each sister pair do not overlap with other sister pairs on the tree. Furthermore, because both lineages in a sister pair have had the same amount of time to accrue substitutions since their common ancestor, it is possible in many cases (see below) to estimate the differ-

Box 3. Why phylogenetic comparative methods are necessary

Most statistical tests assume independence of data points. This independence is often violated for biological data because a single heritable change in a trait can be inherited by many descendants. Treating each of the descendants as independent observations risks counting that single instance of change multiple-times, leading to pseudoreplication. Consider a hypothetical example where rate of molecular evolution is compared between different species, and a correlation analysis is used to detect a relationship between the substitution rate and some biological character (e.g. metabolic rate). The strong correlation between the rate and trait values in Figure Ia might be interpreted as support for a significant association between rates and this trait. But this pattern could have resulted from single, potentially unconnected, changes in both rate and trait (Figure Ib), and this would not be enough information with which to infer a correlation between rate and trait.

In studies of rates, non-independence due to shared history arises in two quite distinct ways. Firstly, with all traits, similarity between lineages is often influenced by relatedness. In Figure I, the rodents are all more similar to each other in metabolic rate than any of them is to a primate, so this will tend to scale with any other character that differs between rodents and primates, including substitution rate. Even aspects of a species' abiotic environment like latitude and temperature will tend to cluster on a phylogeny, as closely related species tend to be found in similar environments. So, phylogenetic non-independence must be addressed whenever any species characteristics that are influenced by relatedness are analysed.

Secondly, some comparative studies of substitution rates include the same substitutions in more than one species' substitution rate estimate, without any correction for this non-independence (discussed further in Ref. [30]). In Figure I, if the rate of molecular evolution in primate species was estimated by comparing a sequence from each species to the human sequence, then the shared history would be counted more than once in the analysis. For example, any molecular changes that occurred in the ancestor of the chimp and human (marked with * in Fig. Ib) would be included in the rate estimate of both gorillas and Orang. Repeatedly sampling the same data point will tend to inflate the apparent significance of any observed relationship between rates and traits.

Phylogenetic comparative methods aim to avoid or correct for both of these problems of phylogenetic non-independence, to prevent analyses being misled by spurious relationships.

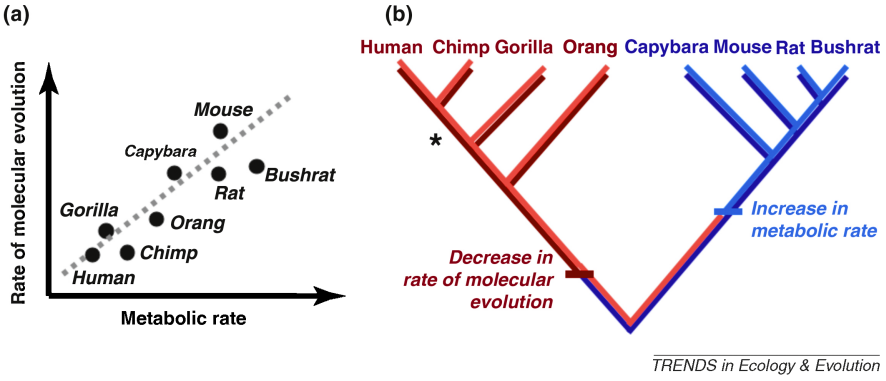

Figure I. A demonstration of phylogenetic non-independence. Because the rodent species have both high metabolic rates and fast rates of molecular evolution, they cluster together on the graph, when compared to primates, which all have lower metabolic rates and slower rates of molecular evolution. Pseudoreplication leads to the false impression of a robust, general relationship.

ence in their rate of molecular evolution simply by comparing their molecular branch lengths, which avoids the need to know their divergence date.

Data from the sister pairs method can be analysed using a variety of statistical tests, the simplest of which is the non-parametric sign test (e.g. [50]). This test simply asks whether the member of the pair that has the higher trait value also tends to have the higher rate value. The sign test makes very few assumptions about the data, because it is based solely on the direction, and not the magnitude, of changes in rates and traits. Because of this, the sign test is unlikely to give false positive results, but has low power

Table 1. Comparative methods appropriate for the study of substitution rates

Rates measured from	Particularly useful when	Example statistical tests and methods	Examples with categorical traits	Example with continuous traits
Sister pairs	Statistical power is less of an issue. Terminal nodes cannot be securely dated. Deep phylogeny is poorly known.	Sign-tests; standard parametric tests of association (e.g., regression or GLMs)	[66,78,79]	[4,5,10]
All terminal branches	Terminal nodes securely dated. A reasonable model of rate/trait evolution can be fitted.	All phylogenetic comparative methods, i.e. methods that account for covariance between the values of a given trait (e.g. Box 4)	[80]	[1]
The complete tree	Ancestral trait values can be securely estimated. A reasonable model of rate/trait evolution can be fitted. Potential correlate is not assignable to the terminal branches (e.g. the number of nodes through which a lineage passes). Node-density effect is unlikely to explain results.	All phylogenetic comparative methods (e.g. Box 4) and multivariate model fitting to complete tree [54].	[54]	[9]

and so can fail to detect more subtle patterns in the data. Data from the sister-pairs method can also be treated as continuous variables and analysed in a parametric (e.g. regression forced through the origin, or Generalised Linear Modelling, GLM) or non-parametric framework (e.g. Wilcoxon Signed Ranks as used in Ref. [12]). These approaches offer considerable increases in power over the sign test, but they also require additional assumptions to be made about the evolutionary processes that have occurred along the lineages in question [43,51]. Standard procedures are available to test these extra assumptions for species' traits [45,51], and with some minor adjustments these tests can also be applied to substitution rates [26] (Table 2). If the data are found to violate any of the assumptions of the statistical tests, each method has an associated 'fix' which can be used before statistical tests of association are applied to the data (see Table 2).

The main drawback of the sister pairs method is that it does not use all of the available information on evolutionary changes in rates and traits, since it ignores changes on unpaired lineages and internal branches of a phylogeny (e.g. deeper nodes of the tree connecting different sister pairs). Because of this, the largest possible number of independent data points generated using this method is equal to half the number of species in the dataset and, depending on the topology, much of the evolutionary change inferable from the data might be neglected.

Methods that use rates estimates from all terminal branches of a tree These methods use dated nodes to convert molecular branch lengths on the terminal branches of a tree into absolute substitution rate estimates, and then test for an association between rates and species traits using phylogenetic comparative methods (e.g. Figure 1b). This approach is attractive because it uses much more of the information about evolutionary changes in rates and traits than the sister pairs approach, and can therefore have more power to detect associations. However, in contrast to the sister pairs approach, this approach always requires that certain assumptions are made about the evolutionary processes that generated the data, and so care must be taken to make sure these assumptions are met. In addition, this approach requires accurate estimation of absolute substitution rates in the terminal branches of the phylogeny, and so the relative dates of divergence of the terminal nodes in the tree need to be known with confidence.

If reliable absolute substitution rate estimates can be obtained for terminal branches (Figure 1b), tests of association between rate and trait can be carried out using comparative methods, which correct for phylogenetic non-independence in the data. The two most commonly used methods in the study of substitution rates are phylogenetically independent contrasts (PIC) [46] and generalised least squares (GLS) regression [52]. The PIC method works like an extension of the sister pairs method: not only are differences in rates and traits calculated between pairs of sister taxa, but rates and traits are also estimated for deeper nodes in the tree, and then compared to one another in a nested fashion [46]. The GLS method is based on an extension of standard regression methods, in which non-independence of data points is estimated using the underlying phylogeny, and this non-independence is then explicitly accounted for when estimating the slope and intercept of the best-fit regression line [52]. The PIC and GLS methods are essentially identical under most standard conditions, but there are some important differences that might make one or other method preferable in certain cases (described in Box 4). These comparative methods are all inherently parametric and so should not be used with non-parametric tests of association (in contrast, e.g. to the sister pairs method).

The correction for the non-independence of data points using either the PIC or GLS methods assumes that the complete phylogeny (with branch lengths) connecting all species in the analysis is known. But for most molecular phylogenies, many branch lengths and topological relationships will be associated with considerable uncertainty. One way around this is to repeat the analysis for a large sample of trees that represent the uncertainty in the data (such as a posterior sample of trees from a Bayesian analysis, e.g. [9,53]). However, it has been shown that rough estimates of branch lengths, though not suitable for rate estimation, can be used to correct for non-independence ([48], Box 4).

Methods that use rate estimates from all branches of a tree There are currently two 'whole-tree' methods that use estimates of substitution rates not only from terminal branches (as in the methods described above), but also from the internal branches of the phylogeny.

The first method uses molecular branch lengths measured from the tip of each terminal branch to the root of the tree (root-to-tip branch lengths), as estimates of substitution rates (e.g. Figure 1c). One use of this approach has been to study links between diversification and molecular evolution [9,33]. In this case, the 'trait' of interest is the number of nodes through which each lineage has passed during its evolutionary history, and the appropriate estimate of substitution rate is therefore the root-to-tip branch length. These data could be analysed by many of the previously described comparative methods (e.g. Box 4),

Table 2. Diagnostic tests of the assumptions of statistical tests that assume independent and identically distributed (i.i.d.) data points (e.g. least-squares regression or GLM), when used with data generated by phylogenetic comparative methods (e.g. sister pairs and PIC)

Assumption being tested	Action if assumption violated	Reference for test
Variance of trait differences is unrelated to their absolute values	Choose appropriate data transformation (e.g. log)	[51]
Variance of trait and rate differences increases linearly with evolutionary time	Transform the branch lengths of the phylogeny, and re-standardise the pairs	[45]
Variance of estimated rate differences is relatively unaffected by stochastic fluctuations in substitution number	Exclude problematic data points	[26]

Box 4. Comparing comparative methods

The two most commonly used phylogenetic comparative methods are generalised least squares regression (GLS; [52]), and phylogenetically independent contrasts (PIC [46]), both of which are implemented in freely-available software (e.g. the 'ape' package in the R environment [74,75]). Both of these methods use a known phylogeny and a model of evolution to correct for correlations between the observed values of a single trait (e.g. those resulting from phylogenetic inertia, Box 3). In many cases the methods are completely equivalent and will give identical results with the same data. However, in their generalised forms they differ in a few important ways.

GLS methods can estimate the extent of the correlations between the observed values of a single trait while simultaneously estimating the regression coefficients between different traits [74,76], allowing for data sets with different amounts of phylogenetic correlation. For a given phylogeny, we can estimate the phylogenetic correlations in different ways, e.g. by applying different transformations to the phylogenetic branch lengths [45], or to the entries of the correlation matrix that the tree implies [76]. Some varieties of GLS do not require a detailed knowledge of the branch lengths [48], while others allow for more complex models of evolution than standard Brownian motion [77].

PIC methods 'transform out' the correlations between the observed values of a single trait, by calculating differences between the weighted averages of the raw species values to create data points that are statistically independent and identically distributed (i.i.d.) [46]. These i.i.d. data points from each trait can then be analysed by standard regression or correlation tests. One benefit of PIC is that different transformations of the branch lengths can be used with each trait, which is helpful when different traits have evolved in different ways [45]. However, PIC methods have one less degree of freedom than GLS methods, and so cannot estimate the intercept of the regression (slopes must be forced through the origin with PIC [45,52]).

PIC and GLS also differ in the ease with which the assumptions of regressions can be tested. GLS has the benefit that each data point represents a single species (rather than a weighted average of multiple species values), and so outlying or suspicious points can be more easily identified. On the other hand many useful diagnostic tests of the regression assumptions can only be used with i.i.d. data points, which are generated by PIC but not GLS (Table 2).

with the sole difference that in addition to correcting for phylogenetic inertia, the comparative method must also account for the counting of the same substitutions in multiple data points (Box 3). The node-density effect is a potential problem when measuring root-to-tip branch lengths, and this has led to some controversy around the use of this method (see above).

A different class of whole-tree method compares the substitution rate estimated for each branch of a molecular phylogeny to the trait value estimated for that branch (Figure 1d). For instance, O'Connor and Mundy [54] introduced a maximum likelihood method in which the relative rate of a proportion of sites in the alignment is assumed to switch between two values, determined by the state of a binary phenotypic trait (whose evolution across the tree is simultaneously inferred). The fit of this model is then compared to that of a simpler model in which the relative rate remains constant.

An important consideration with model-based, whole-tree approaches is that different types of information are used to reconstruct substitution rates versus other traits. Substitution rates are inferred from the gradual accrual of changes along each branch of the phylogeny. Because of this, changes in substitution rate can leave a permanent signature in DNA sequences, such that a transient increase in the rate on an internal branch might be detectable from contemporary data. In contrast, a transient change in a trait such as body size on an internal branch (e.g. an increase in size followed by a decrease) will leave no such signal in contemporary body sizes, so it would not be possible to reconstruct these changes from contemporary data alone, even with a sophisticated model of evolution. This potential mismatch between rate and trait data may mislead tests of association, when rates are estimated from the internal branches of the phylogeny [26,48,55]. Because of this, whole-tree methods will be most powerful when independent data (e.g. fossil remains) are available to reconstruct the ancestral states of the traits. These whole-tree approaches are in their infancy, but are likely to become widely used as they are further generalised.

Conclusion

There is now a wealth of data for exploring the relationships between the rate of genomic evolution and features of species biology such as ecology, life history and environment, and statistical methods developed over the past two decades provide the framework for doing so. As the amount of DNA sequence data continues to grow, the careful application of these comparative methods will doubtless allow many fascinating evolutionary processes and patterns to be revealed.

References

1 Nabholz, B. *et al.* (2008) Strong variations of mitochondrial mutation rate across mammals - the longevity hypothesis. *Mol. Biol. Evol.* 25, 120–130
2 Duffy, S. *et al.* (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276
3 Smith, S.A. and Donoghue, M.J. (2008) Rates of molecular evolution are linked to life history in flowering plants. *Science* 322, 86–89
4 Thomas, J.A. *et al.* (2010) A generation time effect on the rate of molecular evolution in invertebrates. *Mol. Biol. Evol.* 27, 1173–1180
5 Welch, J.J. *et al.* (2008) Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.* 8, 53
6 Mooers, A.O. and Harvey, P.H. (1994) Metabolic rate, generation time, and the rate of molecular evolution in birds. *Mol. Phylogenet. Evol.* 3, 344–350
7 Sarich, V.M. and Wilson, A.C. (1973) Generation time and genomic evolution in primates. *Science* 179, 1144–1147
8 Bromham, L. (2002) Molecular clocks in reptiles: Life history influences rate of molecular evolution. *Mol. Biol. Evol.* 19, 302–309
9 Pagel, M. *et al.* (2006) Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* 314, 119–121
10 Davies, T.J. *et al.* (2004) Environmental energy and evolutionary rates in flowering plants. *Proc. Biol. Sci.* 271, 2195–2200
11 Davies, T.J. and Savolainen, V. (2006) Neutral theory, phylogenies, and the relationship between phenotypic change and evolutionary rates. *Evolution* 60, 476–483
12 Bromham, L. *et al.* (2002) Testing the relationship between morphological and molecular rates of change along phylogenies. *Evolution* 56, 1921–1930
13 Martin, A.P. and Palumbi, S.R. (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U. S. A.* 90, 4087–4091
14 Welch, J.J. and Bromham, L. (2005) Molecular dating when rates vary. *Trends Ecol. Evol.* 20, 320–327
15 Smith, A.B. *et al.* (2006) Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). *Mol. Biol. Evol.* 23, 1832–1851
16 Brown, J.M. *et al.* (2010) When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. *Syst. Biol.* 59, 145–161

17 Ohta, T. (1972) Population size and rate of evolution. *J. Mol. Evol.* 1, 305–314
18 Jones, K.E. *et al.* (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* 90, 2648–12648
19 Swofford, D.L. (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates.
20 Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690
21 Guindon, S. *et al.* (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* 33, W557–559
22 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574
23 Suchard, M.A. and Rambaut, A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25, 1370–1376
24 Simon, C. *et al.* (2006) Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Ann. Rev. Ecol. Evol. S.* 37, 545–579
25 Sullivan, J. and Joyce, P. (2005) Model selection in phylogenetics. *Ann. Rev. Ecol. Evol. S.* 36, 445–466
26 Welch, J.J. and Waxman, D. (2008) Calculating independent contrasts for the comparative study of substitution rates. *J. Theor. Biol.* 251, 667–678
27 Charlesworth, D. (2010) Don't forget the ancestral polymorphisms. *Heredity*, doi:10.1038/hdy.2010.14 [Epub ahead of print]
28 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
29 Ho, S.Y. and Jermiin, L. (2004) Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* 53, 623–637
30 Lanfear, R. *et al.* (2007) Metabolic rate does not calibrate the molecular clock. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15388–15393
31 Phillips, M.J. (2009) Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. *Gene* 441, 132–140
32 Hugall, A.F. and Lee, M.S. (2007) The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution* 61, 2293–2307
33 Webster, A.J. *et al.* (2003) Molecular phylogenies link rates of evolution and speciation. *Science* 301, 478
34 Webster, A.J. *et al.* (2004) Response to comments on "Molecular phylogenies link rates of evolution and speciation". *Science* 303, 173
35 Witt, C.C. and Brumfield, R.T. (2004) Comment on "Molecular phylogenies link rates of evolution and speciation" (I). *Science* 303, 173 author reply 173
36 Brower, A.V. (2004) Comment on "Molecular phylogenies link rates of evolution and speciation" (II). *Science* 303, 173 author reply 173
37 Venditti, C. *et al.* (2006) Detecting the node-density artifact in phylogeny reconstruction. *Syst. Biol.* 55, 637–643
38 Venditti, C. and Pagel, M. (2008) Model misspecification not the node-density artifact. *Evolution* 62, 2125–2126
39 Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19, 101–109
40 Nabholz, B. *et al.* (2009) The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals. *BMC Evol. Biol.* 9, 54
41 Ho, S.Y. (2009) An examination of phylogenetic models of substitution rate variation among lineages. *Biol. Lett.* 5, 421–424
42 Cutler, D.J. (2000) Understanding the overdispersed molecular clock. *Genetics* 154, 1403–1417
43 Harvey, P.H. and Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*, Oxford University Press
44 Freckleton, R.P. (2009) The seven deadly sins of comparative analysis. *J. Evol. Biol.* 22, 1367–1375
45 Garland, T. *et al.* (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41, 18–32
46 Felsenstein, J. (1985) Phylogenies and the comparative method. *Am. Nat.* 125, 1–15
47 Freckleton, R.P. *et al.* (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* 160, 712–726
48 Grafen, A. (1989) The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 326, 119–157
49 Freckleton, R.P. and Harvey, P.H. (2006) Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biol.* 4, e373
50 Lanfear, R. and Bromham, L. (2008) Statistical tests between competing hypotheses of Hox cluster evolution. *Syst. Biol.* 57, 1–11
51 Freckleton, R.P. (2000) Phylogenetic tests of ecological and evolutionary hypotheses: checking for phylogenetic independence. *Funct. Ecol.* 14, 129–134
52 Pagel, M. (1998) Inferring evolutionary processes from phylogenies. *Zool. Scr.* 26, 331–348
53 Huelsenbeck, J.P. and Rannala, B. (2003) Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution* 57, 1237–1247
54 O'Connor, T.D. and Mundy, N.I. (2009) Genotype-phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. *Bioinformatics* 25, i94–100
55 Oakley, T.H. and Cunningham, C.W. (2000) Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution* 54, 397–405
56 Laird, C.D. *et al.* (1969) Rate of fixation of nucleotide substitutions in evolution. *Nature* 224, 149–154
57 Wu, C.I. and Li, W.H. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. U. S. A.* 82, 1741–1745
58 Gillooly, J.F. *et al.* (2005) The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. U. S. A.* 102, 140–145
59 Bromham, L. *et al.* (1996) Determinants of rate variation in mammalian DNA sequence evolution. *J. Mol. Evol.* 43, 610–621
60 Gruber, J. *et al.* (2008) The mitochondrial free radical theory of ageing–where do we stand? *Front. Biosci.* 13, 6554–6579
61 Galtier, N. *et al.* (2009) Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. *Biol. Lett.* 5, 413–416
62 Bromham, L. (2009) Why do species vary in their rate of molecular evolution? *Biol. Lett.* 5, 401–404
63 Chamary, J.V. *et al.* (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7, 98–108
64 Eyre-Walker, A. and Keightley, P.D. (2007) The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618
65 Woolfit, M. (2009) Effective population size and the rate and pattern of nucleotide substitutions. *Biol. Lett.* 5, 417–420
66 Bromham, L. and Leys, R. (2005) Sociality and the rate of molecular evolution. *Mol. Biol. Evol.* 22, 1393–1402
67 Woolfit, M. and Bromham, L. (2003) Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol. Biol. Evol.* 20, 1545–1555
68 Pond, S.L. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679
69 Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591
70 Aris-Brosou, S. and Bielawski, J.P. (2006) Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. *Gene* 378, 58–64
71 Anisimova, M. and Kosiol, C. (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.* 26, 255–271
72 Bierne, N. and Eyre-Walker, A. (2003) The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* 165, 1587–1597
73 Pond, S.K. and Muse, S.V. (2005) Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* 22, 2375–2385
74 Paradis, E. *et al.* (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290
75 R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
76 Freckleton, R.P. *et al.* (2003) Bergmann's rule and body size in mammals. *Am. Nat.* 161, 821–825

77 Martins, E.P. and Hansen, T.F. (1997) Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* 149, 646–667
78 Woolfit, M. and Bromham, L. (2005) Population size and molecular evolution on islands. *Proc. Biol. Sci.* 272, 2277–2282
79 Wright, S. *et al.* (2006) The road from Santa Rosalia: a faster tempo of evolution in tropical climates. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7718–7722
80 Ives, A.R. and Garland, T., Jr (2010) Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* 59, 9–26