

Molecular phylogenetics: state-of-the-art methods for looking into the past

Simon Whelan, Pietro Liò and Nick Goldman

As the amount of molecular sequence data in the public domain grows, so does the range of biological topics that it influences through evolutionary considerations. In recent years, a number of developments have enabled molecular phylogenetic methodology to keep pace. Likelihood-based inferential techniques, although controversial in the past, lie at the heart of these new methods and are producing the promised advances in the understanding of sequence evolution. They allow both a wide variety of phylogenetic inferences from sequence data and robust statistical assessment of all results. It cannot remain acceptable to use outdated data analysis techniques when superior alternatives exist. Here, we discuss the most important and exciting methods currently available to the molecular phylogeneticist.

Only through MOLECULAR PHYLOGENETIC (see Glossary) studies can we understand fully the rapidly accumulating genomic sequence data and information regarding proteins' structure and function. Phylogenetic analyses used to be restricted to descriptive and speculative studies of evolutionary relationships, but recent advances in methodology allow us to recognize and exploit statistical dependencies among sequences sharing common ancestry in broader fields of research¹.

Advances in two areas led to these radically improved methodologies. First, mathematical models of sequence evolution were improved greatly by the incorporation of biological, biochemical and evolutionary knowledge. Second, statistical testing has become an indispensable tool for phylogeneticists, allowing robust evaluation of complex evolutionary hypotheses². The widespread use of accurate models of evolution and statistical tests is necessary to extract the maximum amount of information from molecular sequence data. Many examples in the literature are still employing outdated methods and risk drawing erroneous or weakened conclusions. Here, we review the state of the art in practical molecular phylogenetic analysis.

Modelling evolution

To be both powerful and robust, statistical inference techniques require accurate probabilistic models of the biological processes that generate the data observed. For the phylogenetic analysis of aligned sequences, virtually all methods describe sequence evolution using a model that consists of two components: a PHYLOGENETIC TREE and a description of the way individual sequences evolve by nucleotide or amino acid replacement along the branches of that tree. These replacements are usually described as the products of chance mutation events, and their

occurrence at each sequence site is mathematically modelled by a MARKOV PROCESS. Typically, the same Markov process model is applied to each sequence site. Different models are distinguished by their ASSUMPTIONS OF PARAMETERIZATIONS regarding the average rates of occurrence of all the possible replacements³. It is now well established that more complex models, often describing replacement rates in terms of a variety of biological phenomena, generally give a statistically better fit to observed patterns of sequence evolution⁴⁻⁹, giving more accurate and robust estimates both of phylogeny¹⁰⁻¹² and the statistical confidence in the phylogeny¹³. The development of more accurate models of sequence evolution has been a particularly active and fertile area of research in recent years, in the expectation that improved models will lead to improved phylogenetic inferences and understanding of molecular sequence evolution.

There are two approaches to building models of sequence evolution. One approach is to build models EMPIRICALLY using properties calculated through comparisons of large numbers of observed sequences; for example, simply counting apparent replacements between many closely related sequences. Empirical models result in fixed parameter values, which are estimated only once and then assumed to be applicable to all datasets. This makes them computationally easy to use, but the breadth of the applicability has to be considered carefully because there can be little or no way for these models to be influenced by the data analysed using them. The alternative approach is for models to be built PARAMETRICALLY on the basis of the chemical or biological properties of DNA and amino acids; for example, incorporating a parameter to describe the relative frequency of transition and transversion substitutions in the DNA sequences studied (see below). Parametric models allow the parameter values to be derived from the dataset in each particular analysis. Both methods result in Markov process models, defined by matrices containing the relative rates (i.e. the relative numbers, on average and per unit time) of occurrence of all possible replacements. From these are calculated the probabilities of change from any nucleotide to any other nucleotide (or any amino acid to any other amino acid), including the probability of remaining the same, over any period of evolutionary time (e.g. from one end of a branch to the other) at any site^{3,14}.

S. Whelan
P. Liò
N. Goldman*
University Museum of
Zoology, Dept of Zoology,
University of Cambridge,
Downing Street,
Cambridge, UK CB2 3EJ.
*e-mail: N.Goldman@
zoo.cam.ac.uk

Glossary

Bootstrap: A statistical method by which distributions that are difficult to calculate exactly can be estimated by the repeated creation and analysis of artificial datasets. In the non-parametric bootstrap, these datasets are generated by resampling from the original data, whereas in the parametric bootstrap, the data are simulated according to the hypothesis being tested. The name derives from the near-miraculous way in which the method can 'pull itself up by its bootstraps' and generate statistical distributions from almost nothing.

Empirical and parametric models (parameterization): Most mathematical models of sequence evolution include variables that represent features of the process of evolution, but the numerical values of which are not known *a priori*. These variables are termed parameters of the models. Empirically constructed models take the values of their parameters from pre-computed analyses of large quantities of data, with the particular data under analysis having limited or no influence. By contrast, parametric models do not have pre-specified parameter values. Maximum likelihood can be used to estimate such values from the data under analysis.

Likelihood ratio test (LRT): A powerful form of statistical test in which competing hypotheses (H_0 and H_1) are compared using a statistic based on the ratio of the maximum likelihoods (\hat{l}_0, \hat{l}_1) under each hypothesis; for example, $2\delta = 2\ln(\hat{l}_1/\hat{l}_0) = 2(\ln\hat{l}_1 - \ln\hat{l}_0)$. Results can be expressed in terms of *P*-values, the probability of the statistic being at least as extreme as observed when H_0 is true: low *P*-values (e.g. <0.05) suggest rejection of H_0 in favour of H_1 .

Markov process: A mathematical model of infrequent changes of (discrete) states over time, in which future events occur by chance and depend only on the current state, and not on the history of how that state was reached. In molecular phylogenetics, the states of the process are the possible nucleotides or amino acids present at a given time and position in a sequence and state changes represent mutations in sequences.

Maximum likelihood (ML): The likelihood (l_H) of a hypothesis (H) is equal to the probability of observing the data if that hypothesis were correct. The statistical method of maximum likelihood (ML) chooses amongst hypotheses by selecting the one which maximizes the likelihood; that is, which renders the data the most plausible. In the context of molecular phylogenetics, a model of nucleotide or amino acid replacement permits the calculation of the likelihood for any possible

combination of tree topology and branch lengths. The topology and branch lengths that maximize this likelihood (or, equivalently, its natural logarithm, $\ln l_H$, which is almost invariably used to give a more manageable number) are the ML estimates. Any parameters with values not explicitly specified by the replacement model can be simultaneously estimated, again by selecting the values that maximize the likelihood.

Molecular phylogenetics: The study of phylogenies and processes of evolution by the analysis of DNA or amino acid sequence data.

Phylogenetic tree: The hierarchical relationships among organisms arising through evolution. In his *Origin of Species*⁸, Darwin's only figure uses a sketch of a tree-like structure to describe evolution: from ancestors at the limbs and branches of the tree, through more recent ancestors at its twigs, to contemporary organisms at its buds. Today, these relationships are usually represented by a schematic 'tree' comprising a set of nodes linked together by branches. Terminal nodes (tips or leaves) typically represent known sequences from extant organisms. Internal nodes represent ancestral divergences into two (or more) genetically isolated groups; each internal node is attached to one branch representing evolution from its ancestor, and two (or more) branches representing its descendants. The lengths of the branches in the tree can represent the evolutionary distances that separate the nodes; the tree topology is the information on the order of relationships, without consideration of the branch lengths.

Rate heterogeneity and gamma distribution: Mutation rates vary considerably amongst sites of DNA and amino acid sequences, because of biochemical factors, constraints of the genetic code, selection for gene function, etc. This variation is often modelled using a gamma distribution of rates across sequence sites. The shape of the gamma distribution is controlled by a parameter α , and the distribution's mean and variance are 1 and $1/\alpha$, respectively. Large values of α (particularly $\alpha > 1$) give a bell curve-shaped distribution, suggesting little or no rate heterogeneity; few such examples appear in the literature¹⁹. Small values of α give a reverse - J-shaped distribution, suggesting higher levels of rate heterogeneity along with many sites with low rates of evolution.

Reference

a Darwin, C. (1859) *On the Origin of Species*, John Murray

Most models assume reversibility of this matrix⁷; that is, that the process of evolution would appear the same going backwards in time as it does going forwards (so, for instance, observing an A in one sequence changing to a T in another has the same probability as observing a T changing to an A). No inferences about evolutionary direction can be made when making the assumption of reversibility, unless further information extrinsic to the sequences themselves (for instance, the fossil record¹⁵) is supplied.

Models of DNA substitution

Modelling of DNA evolution has concentrated on the parametric approach. There are three main types of parameters used in these models: (1) base frequency parameters, (2) base exchangeability parameters, and (3) RATE HETEROGENEITY parameters.

The base frequency parameters describe the frequencies of the bases A, C, G and T, averaged over all sequence sites and over the tree. These parameters can be considered to represent constraints on base frequencies due to effects such as overall GC content, and act as weighting factors in a model by making certain bases more likely to arise when substitutions occur.

Base exchangeability parameters describe the relative tendencies of bases to be substituted for one another (up to six parameters, representing the rates

of the changes $A \leftrightarrow C$, $A \leftrightarrow G$, $A \leftrightarrow T$, $C \leftrightarrow G$, $C \leftrightarrow T$ and $G \leftrightarrow T$) discounting base frequency effects. These parameters might be considered to represent a measure of the biochemical similarity of bases: the greater the similarity, the more we expect to see one be substituted for another. For example, transition substitutions (purine \leftrightarrow purine, i.e. $A \leftrightarrow G$, and pyrimidine \leftrightarrow pyrimidine, i.e. $C \leftrightarrow T$) usually occur more often than transversion substitutions (purine \leftrightarrow pyrimidine) in DNA¹⁶. Commonly, transitions are assigned a rate, κ , relative to a rate of 1 for transversions¹⁷, and κ is usually found to be significantly greater than 1.

The most widespread approach to modelling rate heterogeneity amongst sequence sites is to describe each site's rate as a random draw from a GAMMA DISTRIBUTION^{18,19}. The use of a gamma distribution to describe the heterogeneity of rates of evolution amongst sequence sites is widely recognized to be an important factor in the fitting of models to data^{8,19}. A gamma distribution can be used in combination with base frequency and exchangeability parameters; models incorporating a gamma distribution are often denoted by the suffix '+Γ'. Other methods for describing rate variation include assigning specific rates of substitution to different parts of the sequence²⁰ (e.g. to the three codon positions of protein coding sequences or to different genes or domains), or designating a proportion of sites as invariable²¹.

Fig. 1. Relationships among some standard models of nucleotide evolution. Six standard models of nucleotide evolution [JC (Ref. 75), FEL (Ref. 62), K2P (Ref. 17), HKY (Ref. 21), the most general reversible model REV (Ref. 7) and REV + Γ] are presented in a flowchart showing relationships between them. For each model, we show the matrix of substitutions rates between nucleotides (represented by a bubble plot where the area of each bubble indicates the corresponding rate), a partial representation of a hominoid phylogeny as inferred by that model from a mitochondrial sequence dataset¹⁶, and the maximum log-likelihood value ($\ln \hat{l}$) obtained. For the REV + Γ model we also show the gamma distribution of rates among sites described by the inferred parameter value $\alpha = 0.28$. The reverse-J shape of the graph indicates that the majority of sites have low rates of evolution, with some sites having high rates of evolution. The JC model assumes that all nucleotide substitutions occur at equal rates. The models become more advanced moving down the figure, as illustrated in the bubble plots by their increasing flexibility in estimating relative replacement rates and as reflected by increasing log-likelihoods. Note how the inferred maximum likelihood phylogeny changes significantly as the models become more advanced (compare JC with K2P); inferred branch lengths also tend to increase (compare REV to REV + Γ). Arrows show where models are nested within each other; that is, where the first model is a simpler form of the next. For example, the JC model is nested within the K2P model (it is a special case arising when κ is fixed equal to 1), but the K2P model is not nested with the FEL model.

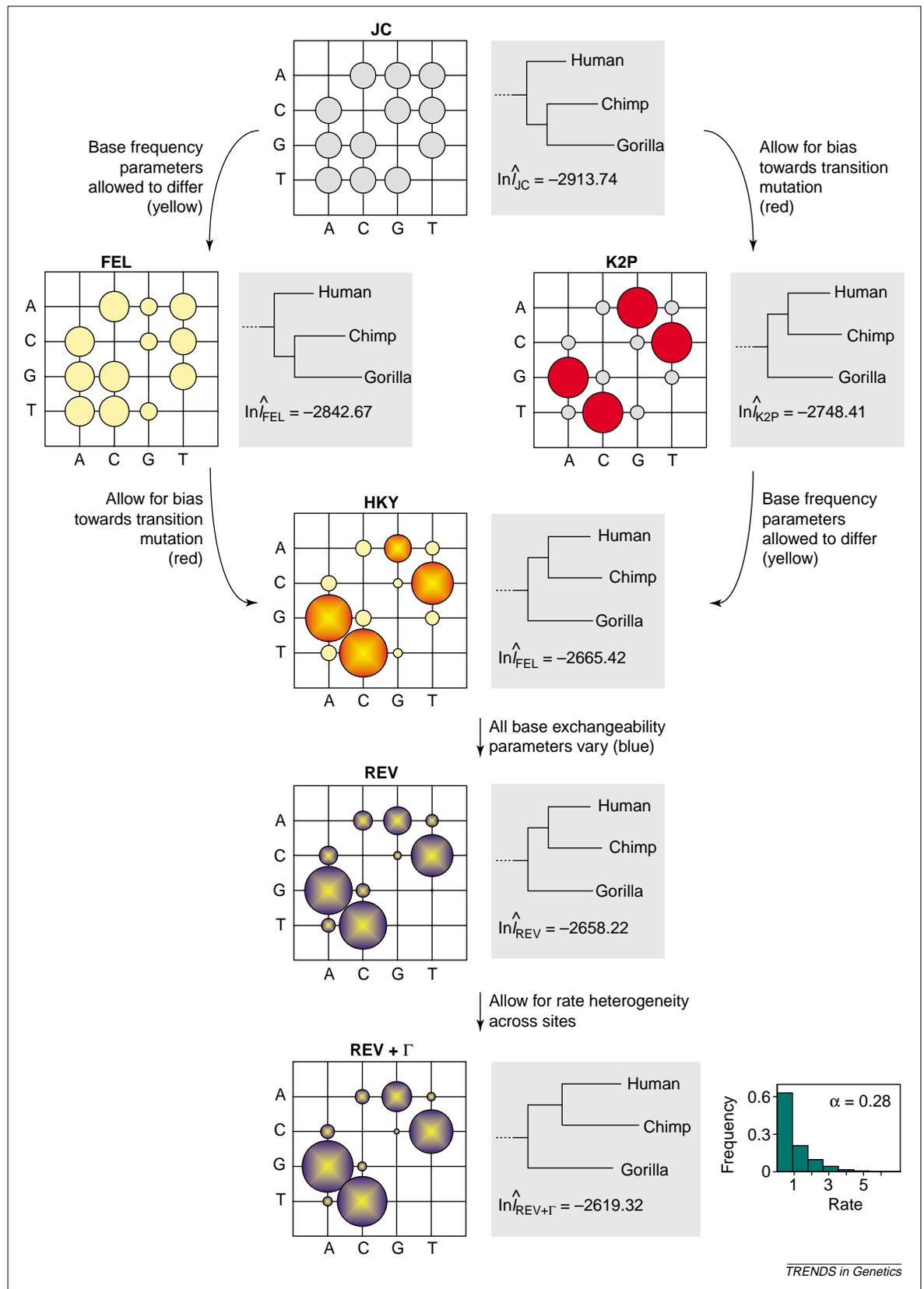


Figure 1 illustrates the relationships among some widely used models of nucleotide substitution, and the differences they can make to the inferences drawn using them. Simpler models do not take account of all of the information from sequences. When used in

phylogenetic inference (see below), they can lead to incorrectly inferred trees and underestimated branch lengths. More advanced and successful models use at least base frequency parameters, the transition/transversion bias parameter κ (as

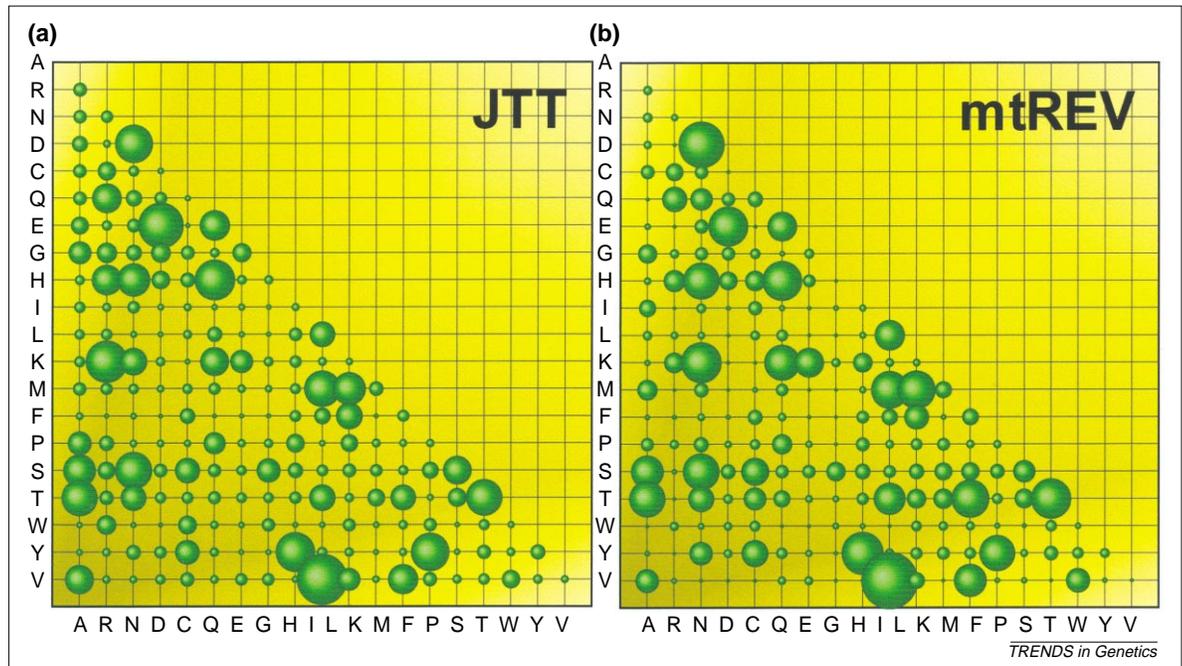


Fig. 2. Exchangeability parameters for two models of amino acid replacement. Exchangeability parameters from two common empirical models of amino acid sequence evolution are presented. The parameter value for each amino acid pair is indicated by the areas of the bubbles, and discounts the effects of amino acid frequencies. (a) The JTT model²⁵ derived from a wide variety of globular proteins. (b) The mtREV model⁹ derived from mammalian mitochondrial genes that encode various transmembrane proteins.

with the HKY model²¹) or the full generality of six exchangeability parameters (as with the REV model⁷), and often the '+ Γ ' option. Statistical tests (see below) almost invariably indicate that these models provide a significantly better description of the evolutionary process. This, in turn, generally leads to more reliable phylogenetic estimates, and models such as these should be used whenever possible.

Models of amino acid replacement

In contrast to DNA substitution models, the modelling of amino acid replacement has concentrated on the empirical approach. Although analogues of the simplest nucleotide substitution models have been used²², these are now effectively abandoned in favour of more advanced models. Early models for amino acid sequences of globular proteins were those of Dayhoff and collaborators^{23,24} and an updated version (JTT) by Jones *et al.*²⁵ These were derived simply by counting observed amino acid replacements in large sequence databases; only very closely related sequences were considered, to reduce the frequency with which observed replacements (e.g. A→S), were in fact the result of a set of successive unobserved replacements (e.g. A→R→S). The amino acid frequency and exchangeability parameters were computed from these counts. More recently, MAXIMUM LIKELIHOOD methods, which make precise allowance for successive replacements and for the phylogenetic relationships among sequences, have been used to derive models specifically

applicable to proteins encoded in mitochondrial^{4,9}, chloroplast⁵ and nuclear genomes²⁶.

Typically, the phylogenetic analysis of particular amino acid sequences will use frequency parameter values estimated from those data in conjunction with the exchangeability parameters from one of the standard models described above²⁷. This forces the amino acids frequencies of the resulting hybrid model to match those of the observed data (instead of those of the database from which the exchangeability parameters were originally estimated), while still incorporating exchangeability information derived from a broad range of sequences. Such applications are usually denoted by a '+F' suffix. Gamma distributions to describe rate heterogeneity are also often useful. So, for example, a phylogenetic analysis employing the model denoted JTT + F + Γ uses the exchangeability parameters of the JTT model, the amino acid frequencies observed in the sequences under analysis and a gamma distribution to represent among site rate heterogeneity. As with nucleotide substitution models, the differences between the available models describing amino acid sequence evolution are considerable (Fig. 2). It is important that an appropriate model be selected, as this can be crucial to phylogenetic inference.

Complex models of evolution

Evolutionary models explicitly describing selection or structure consistently give significantly improved descriptions of the evolution of protein sequences and are especially valuable in giving new insights into the processes of molecular evolution²⁸. Such models, although more complex and sometimes hindered by additional computational difficulties, are becoming more important. Codon-based models have been developed recently that describe the evolution of coding sequences in terms of both DNA substitutions and the selective forces acting on the protein product^{29–31}. For

example, by studying the relationships between rates of synonymous (amino acid conserving) and nonsynonymous (amino acid altering) DNA substitutions, these models have been used successfully to detect where and when positive selection was important; that is, at which positions in genes^{29,32,33} and in which lineages of phylogenies³⁴. In contrast to earlier methods for studying selection at the molecular level, methods based on these models are able to detect selection without *a priori* knowledge of protein structure or function, and in cases where only some sites are evolving under selective pressures³⁵.

Other models have attempted to associate the heterogeneity of patterns and rates of evolution among sites with the structural organization of RNA or proteins. Models accommodating RNA secondary structural elements have used 16 states to represent all the possible base pairings in stem regions and four states to model loops^{36,37}. Proteins' structural and functional properties have been incorporated into models of amino acid sequence evolution by exploiting knowledge of different patterns of amino acid replacement in different structural contexts^{38–40}, allied with mathematical models of the typical organization of these contexts within globular^{38,41} and transmembrane⁴² protein structures. The simultaneous consideration of phylogeny and protein structure allows information about each to improve inference of the other⁴³.

Inferential methodology

All the models of sequence evolution described above can be used to estimate the phylogenetic tree that generated the observed sequences. Ideally, the inference method used will extract the maximum amount of information available in the sequence data, will combine this with prior knowledge of patterns of sequence evolution (encapsulated in the evolutionary models), and will deal with model parameters (e.g. the transition/transversion bias κ) whose values are not known *a priori*. The three major inference methods for inferring molecular phylogenies that have dominated the literature, maximum likelihood, maximum parsimony and pairwise distances (fuller details of which are given, for example, by Swofford *et al.*¹⁴), satisfy these criteria to varying degrees.

Maximum likelihood inference

The most statistically robust method to achieve these aims is to consider the phylogenetic inference problem in a likelihood framework⁴⁴. The method of maximum likelihood (ML) is one of the standard tools of statistics. In molecular phylogenetics, the ML tree is the one that renders the observed sequences the most plausible, given the chosen model of sequence evolution. It permits the inference of phylogenetic trees using complex evolutionary models – including the ability to estimate model parameters and so make inferences simultaneously about the patterns and processes of evolution – and provides the means for comparing competing trees and models.

A great attraction of the likelihood approach in phylogenetics is the existence of a wealth of powerful statistical theory; for example, the ability to perform robust statistical hypothesis tests (see below) and the knowledge that ML phylogenetic estimates are statistically consistent (given enough data and an adequate model, ML will always give the correct tree topology^{45,46}). The likelihood framework also makes possible modern statistical techniques such as Markov chain Monte Carlo and Bayesian inference⁴⁷. These methods, which are beyond the scope of this review, are just beginning to have a significant impact on phylogenetics^{48–50}. These strong statistical foundations suggest that likelihood techniques are the most powerful for phylogeny reconstruction and for understanding sequence evolution.

ML phylogenetic inference suffers from the fact that each possible tree topology should be assessed individually, and, when examining large numbers of sequences, the number of possible tree topologies is huge⁵¹. Heuristic searches for the ML tree are widely used, but give no guarantee of finding the optimal tree¹⁴. Using recent versions of software and moderately powerful personal computers, ML estimation can be practical for hundreds of sequences⁵². When computational constraints make the use of likelihood methods impractical, other inferential approaches can be used.

Other methods of phylogenetic inference

The two other popular approaches to phylogenetic inference are pairwise distance methods and parsimony. Distance methods use the same models of evolution as ML to estimate the evolutionary distance between each pair of sequences from the set under analysis, and then to fit a phylogenetic tree to those distances. The distances will usually be ML estimates for each pair (considered independently of the other sequences), but the set of all pairwise distances will not be compatible with any tree and a best-fitting phylogeny is derived using non-ML methods. Disadvantages of distance methods include the inevitable loss of evolutionary information when a sequence alignment is converted to pairwise distances⁵³, and the inability to deal with models containing parameters for which the values are not known *a priori* (e.g. κ above).

Maximum parsimony selects the tree or trees that require the fewest evolutionary changes. If the number of changes per sequence position is relatively small, then maximum parsimony approximates ML and its estimates of tree topology will be similar to those of ML estimation^{54,55}. As more-divergent sequences are analysed, the degree of homoplasy (i.e. parallel, convergent, reversed or superimposed changes) increases. The true evolutionary tree becomes less likely to be the one with the least number of changes, and parsimony fails, as it has no adequate means to deal with this. Furthermore, when the true tree has short internal branches and

long terminal branches, a phenomenon can occur whereby the long branches appear to attract one another and can be erroneously inferred to be too closely related. Combinations of conditions when this occurs are often called the 'Felsenstein zone', and parsimony is particularly affected by this problem⁵⁶ because of its inability to deal with homoplasy. In the Felsenstein zone, parsimony becomes increasingly certain of the wrong tree; a property referred to as inconsistency. In addition, parsimony lacks an explicit model of evolution⁵⁴. Although originally seen by some as a strength of parsimony⁵⁷, this has now become a limiting factor preventing fruitful feed-forward between phylogenetic analysis and investigation of the biological implications of different models of evolution.

Simulation studies show that ML methods generally outperform distance and parsimony methods over a broad range of realistic conditions¹⁰⁻¹². Whereas research into ML techniques has concentrated for some years on improving the models of sequence evolution used in phylogenetic analyses, recent developments in distance and parsimony methodology have concentrated on elucidating the relationships of these methods to ML inference, and exploiting this understanding to adapt the methods so that they perform more like ML methods^{55,58-61}.

Statistical testing in phylogenetics

In the past decade, one of the most important topics in evolutionary sequence analysis was the development of methods for the statistical testing of phylogenetic hypotheses. These advances are available almost exclusively within the likelihood framework. They permit assessment of which model provides the best fit for a given dataset – vital for the selection of the optimal model with which to perform phylogenetic inference. Additionally, the rejection of simpler models in favour of those that incorporate additional biological factors gives weight to arguments that those factors have a significant role in sequence evolution. Statistical tests in phylogenetics also permit the assessment of the degree of confidence we have in any given tree topology being the true topology; in summary, they are responsible for the mutual feed-forward between our abilities to estimate better trees and to create more realistic models of evolution.

Statistical tests of models

Model comparisons

The likelihood framework permits estimation of parameter values and their standard errors from the observed data, with no need for any *a priori* knowledge⁸. For example, a transition/transversion bias estimated as $\kappa = 2.3 \pm 0.16$ effectively excludes the possibility that there is no such bias ($\kappa = 1$), whereas $\kappa = 2.3 \pm 1.6$ does not.

Comparisons of two competing models are also possible, using LIKELIHOOD RATIO TESTS^{6,8,62} (LRTs; Fig. 3). Competing models are compared (using their

maximized likelihoods) with a statistic, 2δ , that measures, in effect, how much better an explanation of the data the alternative model gives. To perform a significance test, the distribution of values of 2δ expected under the simpler hypothesis is required; if the observed value of 2δ is too great to be consistent with this distribution (e.g. has a P -value < 0.05), the simpler model is rejected in favour of the more complex model.

When the two models of evolution being compared are nested (that is, the simpler model is a special case of the more complex model obtained by constraining certain free parameters to take particular values), then the required distribution for 2δ is usually a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models^{6,8,62}. A notable exception is the recent finding that a different distribution is required when testing for the existence of gamma-distributed rate heterogeneity^{63,64}. When the models being compared are not nested, as can often be the case for more complex models of sequence evolution, the required distribution of 2δ can be estimated by a procedure known as Monte Carlo simulation or parametric BOOTSTRAPPING^{2,6}.

Testing the fit of one model

A test also exists for assessing whether one particular model is a statistically adequate description of the evolution of a set of sequences⁶ (Fig. 3). This test almost invariably indicates that current models of sequence evolution are not explaining the evolutionary patterns in the data fully. In the very few cases where the test has indicated a model to be adequate, the model has been relatively complex and the sequences studied are believed to be under little selection⁸. This suggests that selective effects are amongst the most important factors not yet fully incorporated into models of sequence evolution.

Statistical tests of tree topologies

Non-parametric bootstrapping of phylogeny

In many applications, the primary interest is in the topology of the inferred evolutionary tree. As with estimates of model parameters, a single point-estimate is of little value without some measure of the confidence we can place in it. A popular way of assessing the robustness of a tree is by the method of non-parametric bootstrapping^{14,65} (Fig. 4). Comparisons of an inferred tree with the set of bootstrap replicate trees, typically in the form of tabulation of the proportion of the bootstrap replicates in which each branch from the inferred tree occurs, can give indications of the robustness of the inferred tree.

One difficulty with this analysis is the precise interpretation of what these values represent. Usually they are taken as a representation of statistical confidence in the monophyly of groups of sequences, internally related in any manner. For

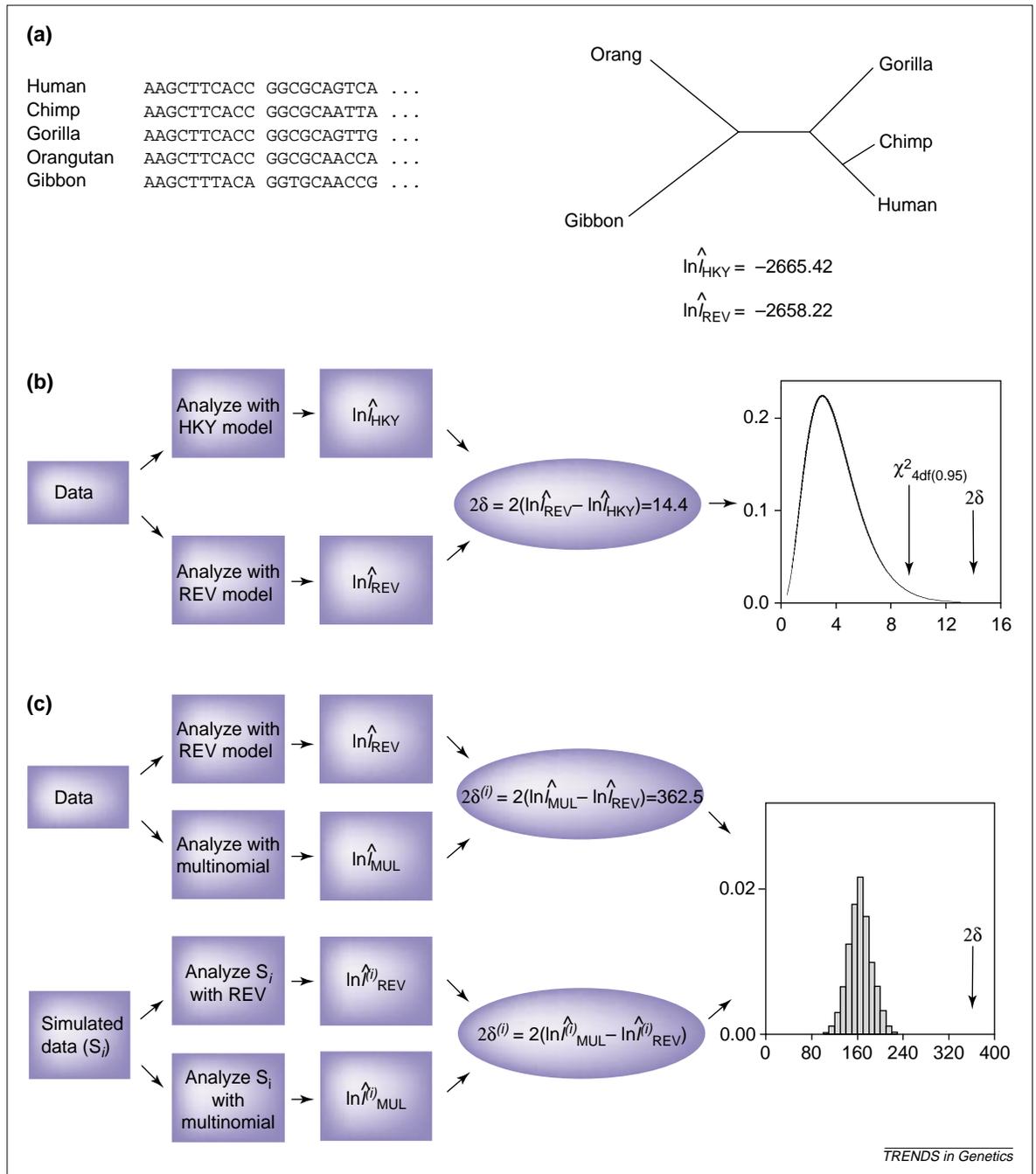


Fig. 3. Statistical tests of models. (a) Part of the mitochondrial sequence dataset¹⁶ used in Fig. 1, and the maximum likelihood phylogenetic tree and likelihood values from the HKY (Ref. 21) and REV (Ref. 7) models. (b) The statistical test to compare these models of nucleotide substitution, in which the likelihood ratio statistic 2δ is compared with a χ^2_4 distribution^{9,62}. The observed value of 2δ , 14.4, has a *P*-value considerably less than 0.05, and the HKY model is rejected in favour of the REV model. (c) The test of the adequacy of the REV model. The test statistic is derived from a comparison of the REV model and a multinomial model that identifies the maximum possible likelihood attainable under any model. The test distribution is estimated by parametric bootstrapping, in which simulated datasets S_i (generated using the maximum likelihood phylogeny and substitution model parameters estimated with the REV model) and are subjected to the same analysis as the original data^{2,6}. Comparison of the test statistic and the distribution of values obtained from simulated data indicates that the observed value 2δ is far in excess of what is expected if the REV model were accurate, and we can conclude that a more complex evolutionary model is necessary to describe the patterns of evolution of these sequences fully.

example, the appearance of a branch in $\geq 95\%$ of the bootstrap replicate trees is often taken as indicating a significant level of confidence in the monophyly of the sequences descended from that branch. However, it becomes very difficult to assess overall confidence in multiple regions or the entirety of a tree because of the unclear relationships of bootstrap values assigned to different groups. For example, if two groups are each given bootstrap values of 90% then, assuming they are independent, the chance that both groups are correct is $(90/100)^2 = 81\%$. As more groups are considered simultaneously, these values decrease further and for large trees the overall values very quickly become meaningless. Bootstrap values may be most suitable for examining small parts of the tree, for example one or two key branches.

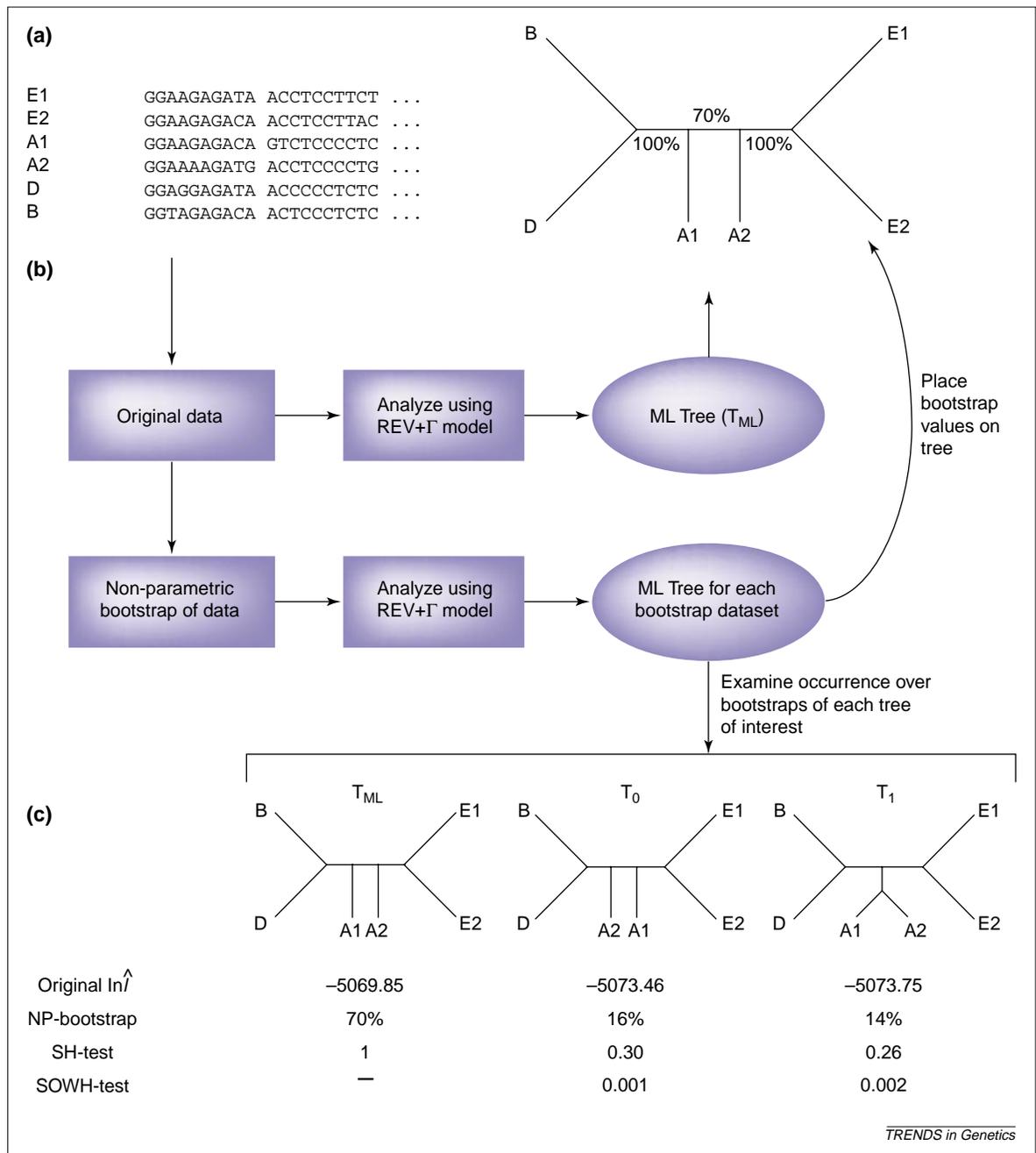


Fig. 4. Statistical tests of tree topologies. (a) Part of a dataset of six HIV-1 nucleotide sequences, from HIV-1 subtypes A (two examples), B, D and E (two examples)⁶⁸, and the maximum likelihood (ML) phylogenetic inference under the REV+ Γ model. Note that the ML phylogeny (T_{ML}) differs from the conventional tree (T_c), which would group the two subtype A sequences together. (b) A non-parametric bootstrap analysis^{14,65} of confidence in T_{ML} : analysis of many bootstrapped datasets allows calculation of the proportion of replicates in which branches appearing in T_{ML} also arise in the bootstrap trees – these values are indicated in (a). Note that the central branch does not receive a statistically significant bootstrap proportion, indicating that there is some uncertainty about the position of the subtype A sequences. (c) The likelihoods assigned to T_{ML} , T_1 and another plausible tree T_0 , and the proportion of the time these trees are inferred from non-parametric bootstrap datasets. Note that T_0 and T_1 are each recovered a considerable proportion of the time. Also shown are the P -values obtained from the non-parametric SH-test⁶⁹, which suggest that none of the three trees shown is rejected as explanations of the observed data, and from the parametric SOWH-test^{14,68}, which rejects both T_1 and T_0 in favour of T_{ML} . As discussed in the text and by Goldman *et al.*⁶⁸, it is not yet fully clear why the results of the SH- and SOWH-tests can be so different.

Statistical tests of competing phylogenetic tree hypotheses

LRTs exist for comparing hypotheses concerning entire phylogenies (Fig. 4). The competing hypotheses are now trees and the test statistic is again based on the maximal likelihood scores under competing hypotheses. There are no known theoretical results to predict the appropriate statistical distributions needed to perform significance tests on trees to choose between them. Such tests therefore first became practical approximately ten years ago, when Kishino and Hasegawa^{66,67} devised a non-parametric bootstrapping procedure which permitted estimation of the necessary distribution in the case that the competing hypotheses each consisted of specific phylogenetic tree topologies, chosen *a priori* (i.e. before the data were analysed). The

Box 1. Software for molecular phylogenetics

This list is only a small selection of software available for molecular phylogenetics, describing programs most closely related to the techniques described in this review. For a much more complete list, see: <http://evolution.genetics.washington.edu/phylip/software.html>

EDIBLE (Experimental Design and Information by Likelihood Exploration^a)

A software tool for experimental design in phylogenetics, using Fisher information to estimate the relative benefits of adding additional taxa, increasing sequence length, or looking at sequences which have evolved at different evolutionary rates. <http://www.zoo.cam.ac.uk/zoostaff/goldman/info/edible.html>

MOLPHY (Molecular Phylogenetics^b)

A package of free programs performing ML phylogenetic inference from DNA and amino acid sequences, including tree searching capabilities. <http://www.ism.ac.jp/software/ismlib/softother.e.html>

PAML (Phylogenetic Analysis by Maximum Likelihood^c)

A free package of programs intended mainly for phylogenetic analysis and evolutionary model comparison. The package includes a wide variety of advanced models, including DNA- and amino acid-based models, and codon-based models that can be used to detect positive selection. Many of the programs are also capable of modelling heterogeneity of evolutionary rates amongst sequence sites using gamma distributions, and of evolutionary dynamics of different sequence regions (e.g. for concatenated gene sequences). <http://abacus.gene.ucl.ac.uk/software/paml.html>

PAUP* [Phylogenetic Inference Using Parsimony (*and other methods^d)]

A commercially available program that, despite its name, implements a wide variety of methods for phylogenetic inference. Maximum likelihood analyses may be performed on DNA data using a variety of models. PAUP* includes a comprehensive set of methods, exact and heuristic, for searching for optimal trees. <http://www.sinauer.com/Titles/frswofford.htm>

PHYLIP (Phylogenetic Inference Package^e)

A large package of free programs for phylogenetic inference using a wide range of methods, including pairwise distance and

parsimony techniques as well as maximum likelihood. The maximum likelihood programs allow a few relatively simple models and have good tree searching capabilities. <http://evolution.genetics.washington.edu/phylip.html>

Seq-Gen (Sequence Generator^f) and PSeq-Gen (Protein Sequence Generator^g)

Seq-Gen and PSeq-Gen are free programs that simulate the evolution of nucleotide and amino acid sequences along a specified phylogeny. This is useful for performing parametric bootstrap tests. Many standard models of nucleotide and amino acid evolution are implemented.

<http://evolve.zoo.ox.ac.uk/software/Seq-Gen/main.html>

<http://evolve.zoo.ox.ac.uk/software/PSeq-Gen/main.html>

Statistical tests of topologies

The program 'shtests' is available for performing the non-parametric bootstrap test of Shimodaira and Hasegawa^h, via: <http://evolve.zoo.ox.ac.uk/software/SHTests/main.html>
Notes on using other software to perform parametric bootstrap tests of topologies^{i,j} are available via: <http://www.zoo.cam.ac.uk/zoostaff/goldman/tests>

References

- a Massingham, T. and Goldman, N. (2000) EDIBLE: experimental design and information calculations in phylogenetics. *Bioinformatics* 16, 294–295
- b Adachi, J. and Hasegawa, M. (1995) *MOLPHY: Programs for Molecular Phylogenetics Ver. 2.3*, Institute of Statistical Mathematics, Tokyo
- c Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556
- d Swofford, D.L. (1999) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Sinauer Association
- e Felsenstein, J. (1993) *PHYLIP (Phylogeny Inference Package), ver. 3.5c*, Dept of Genetics, University of Washington
- f Rambaut, A. and Grassly, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *CABIOS* 13, 235–238
- g Grassly, N.C. *et al.* (1997) PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. *CABIOS* 13, 559–560
- h Shimodaira, H. and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116
- i Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M., *et al.*, eds), pp. 407–514, Sinauer Associates
- j Goldman, N. *et al.* (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49, 652–670

'Kishino–Hasegawa test' or 'KH-test' gained rapid acceptance and is implemented in many phylogenetic software packages.

Unfortunately, the constraint that the topologies being tested should be selected *a priori* became overlooked. The vast majority of applications in recent years have been to the case where one hypothesis topology was specified *a priori*, but the alternative hypotheses topology was taken (*a posteriori*) from a ML analysis of the same data on which the test was to be performed. The intention would often be to use novel data to investigate the acceptability (or otherwise) of a previously published phylogeny. However, as recently highlighted in two

important papers^{68,69}, the use of the same data both to derive an alternative hypothesis topology and to perform the statistical test invalidates the KH-test. In the majority of interesting cases, only a complete reanalysis of the data can confirm (or refute) the conclusions originally drawn from the test⁶⁸.

Fortunately, replacement tests are available (Fig. 4). Shimodaira and Hasegawa⁶⁹ have described a non-parametric bootstrap test that directly succeeds the KH-test, considering all possible topologies and making the proper allowance for their comparison with the ML topology derived from the same data. Parametric bootstrap tests are also available if a parametric approach is preferred,

either for the situation described above^{14,68} or when a hypothesis is not one particular topology but instead comprises a constraint on topology; for example, the monophyly of a pre-specified group of sequences⁷⁰. The precise relationship between the non-parametric and parametric versions of the new tests is not yet clear⁶⁸: it appears that the non-parametric test is very conservative (i.e. unwilling to reject topologies as untrue).

Increasing the robustness of a tree

The best possible phylogenetic estimates will arise from using robust inference methods allied with accurate evolutionary models. However, after statistical assessment of the results it could still be necessary to attempt to improve the quality of inferences drawn. The two most obvious ways of increasing the accuracy of a phylogenetic inference are to include more sequences in the data or to increase the length of the sequences used. Until recently, the likely effects of these approaches had not been well characterized. One recent attempt at quantifying these effects⁷¹ makes use of the statistical theory of the information content of data, again relying on likelihood methods. This study shows that adding more sequences to an analysis does not increase the amount of information relating to different parts of the tree uniformly over that tree, whereas the use of longer sequences results in a linear increase in information over the whole of the tree. Such methods allow questions of experimental design in phylogenetic analysis to be answered; for example, regarding numbers and lengths of sequences to analyse and the identification of genes with optimal rates of evolution for phylogenetic inference^{71–73}.

It is now often the case that several genes are available for all or some of the species a researcher wishes to study. A potentially powerful approach is to analyse the sequences as a concatenated whole or 'meta-sequence'. The simplest analysis is then to assume that all the genes have the same patterns and rates of evolution²⁷. This naive method should only be used when there is substantial evidence of a consistent evolutionary pattern across all the genes, which can be assessed by statistical tests of different models as described above – otherwise, differences amongst genes' replacement patterns or rates can lead to biased results. More advanced analyses of concatenated sequences are possible, which allow for

heterogeneity of evolutionary patterns among the genes studied²⁰. This heterogeneity might be as simple as modelling all genes as having the same patterns of replacement but different rates throughout their common tree, or as complex as allowing each gene to evolve with different replacement patterns, and with different rates of replacement in all branches of the genes' trees (which now have a common topology but unrelated branch lengths from one gene to the next⁷⁴). LRTs again provide a straightforward means of deciding which topologies best fit the data under analysis. The approach of probabilistic modelling, likelihood analysis and statistical hypothesis testing again allows the selection and use of the most powerful methods for extracting evolutionary information from multiple gene sequences.

Conclusion

Molecular evolutionary studies are central to a huge range of biological areas; this is increasingly true as sequence databases grow (and include numerous whole genomes and proteomes). The phylogenetic methodology required for these studies has progressed greatly in the past few years. Maximum likelihood methods permit the application of mathematical models that incorporate our prior knowledge of typical patterns of sequence evolution accumulated over more than 30 years, resulting in more powerful inferences. At present, only likelihood-based methods are able simultaneously to make inferences about the processes of sequence evolution and to use models of those processes to make inferences about evolutionary relationships. Furthermore, they use a complete statistical methodology that permits hypothesis tests, enabling validation of the results at all stages: from the values of parameters in evolutionary models, through the comparison of competing models describing the biological factors most important in sequence evolution, to the testing of hypotheses of evolutionary relationship. Computer programs for the robust statistical evolutionary analysis of molecular sequence data are widely available (Box 1). It is no longer acceptable to rely on evolutionary estimates without concern either for possible systematic biases, caused by naïve and inadequate models of evolution or unsatisfactory estimation procedures, or for measures of confidence in results.

References

- 1 Harvey, P.H. *et al.* eds, (1996) *New Uses for New Phylogenies*, Oxford University Press
- 2 Huelsenbeck, J.P. and Rannala, B. (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276, 227–232
- 3 Liö, P. and Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Res.* 8, 1233–1244
- 4 Adachi, J. and Hasegawa, M. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42, 459–468
- 5 Adachi, J. *et al.* (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* 50, 348–358
- 6 Goldman, N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36, 182–198
- 7 Yang, Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39, 105–111
- 8 Yang, Z. *et al.* (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324
- 9 Yang, Z. *et al.* (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* 15, 1600–1611
- 10 Gaut, B.S. and Lewis, P.O. (1995) Success of maximum-likelihood phylogeny inference in the 4-taxon case. *Mol. Biol. Evol.* 12, 152–162
- 11 Huelsenbeck, J.P. (1995) Performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48
- 12 Kuhner, M.K. and Felsenstein, J. (1994) Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468. See also Erratum. *Mol. Biol. Evol.* 12, 525 (1995)

- 13 Yang, Z. *et al.* (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* 44, 384–399
- 14 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer Associates
- 15 Page, R.D.M. and Holmes, E. (1998) *Molecular Evolution*, Blackwell Science
- 16 Brown, W.M. *et al.* (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* 18, 225–239
- 17 Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120
- 18 Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39, 306–314
- 19 Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11, 367–372
- 20 Yang, Z. (1996) Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596
- 21 Hasegawa, M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174
- 22 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press
- 23 Dayhoff, M.O. *et al.* (1972) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Vol. 5) (Dayhoff, M.O., ed.), pp. 89–99, National Biomedical Research Foundation
- 24 Dayhoff, M.O. *et al.* (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–352, National Biomedical Research Foundation
- 25 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8, 275–282
- 26 Whelan, S. and Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol. Biol. Evol.* (in press)
- 27 Cao, Y. *et al.* (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.* 39, 519–527
- 28 Thorne, J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* 10, 602–605
- 29 Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936
- 30 Yang, Z. and Nielsen, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409–418
- 31 Yang, Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449
- 32 Bishop, J.G. *et al.* (2000) Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5322–5327
- 33 Zanotto, P.M. *et al.* (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* 153, 1077–1089
- 34 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573
- 35 Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503
- 36 Rzhetsky, A. (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics* 141, 771–783
- 37 Schöniger, M. and von Haeseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylog. Evol.* 3, 240–247
- 38 Goldman, N. *et al.* (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149, 445–458
- 39 Koshi, J.M. and Goldstein, R.A. (1995) Context-dependent optimal substitution matrices. *Protein Eng.* 8, 641–645
- 40 Topham, C.M. *et al.* (1993) Fragment ranking in modelling protein structure: conformationally constrained substitution tables. *J. Mol. Biol.* 229, 194–220
- 41 Thorne, J.L. *et al.* (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* 13, 666–673
- 42 Lio, P. and Goldman, N. (1999) Using protein structural information in evolutionary inference: transmembrane proteins. *Mol. Biol. Evol.* 16, 1696–1710
- 43 Goldman, N. *et al.* (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* 263, 196–208
- 44 Edwards, A.W.F. (1972) *Likelihood*, Cambridge University Press
- 45 Chang, J.T. (1996) Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137, 51–73
- 46 Rogers, J.S. (1997) On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 46, 354–357
- 47 Shoemaker, J.S. *et al.* (1999) Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet.* 15, 354–358
- 48 Larget, B. and Simon, D. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750–759
- 49 Mau, B. *et al.* (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55, 1–12
- 50 Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: Markov chain Monte Carlo methods. *Mol. Biol. Evol.* 14, 717–724
- 51 Felsenstein, J. (1978) The number of evolutionary trees. *Syst. Zool.* 27, 27–33
- 52 Yang, Z. (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.* 51, 423–432
- 53 Steel, M.A. *et al.* (1988) Loss of information in genetic distances. *Nature* 336, 118
- 54 Goldman, N. (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. *Syst. Zool.* 39, 345–361
- 55 Steel, M. and Penny, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17, 839–850
- 56 Huelsenbeck, J.P. (1997) Is the Felsenstein zone a fly trap? *Syst. Biol.* 46, 69–74
- 57 Platnick, N.I. (1985) Philosophy and the transformation of cladistics revisited. *Cladistics* 1, 87–94
- 58 Bruno, W.J. *et al.* (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* 17, 189–197
- 59 Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685–695
- 60 Ota, S. and Li, W.-H. (2000) NJML: a hybrid method for the neighbor-joining and maximum-likelihood methods. *Mol. Biol. Evol.* 17, 1401–1409
- 61 Willson, S.J. (1999) A higher order parsimony method to reduce long-branch attraction. *Mol. Biol. Evol.* 16, 694–705
- 62 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376
- 63 Goldman, N. and Whelan, S. (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 17, 975–978
- 64 Ota, R. *et al.* (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* 17, 798–804
- 65 Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791
- 66 Hasegawa, M. and Kishino, H. (1989) Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* 43, 672–677
- 67 Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29, 170–179
- 68 Goldman, N. *et al.* (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49, 652–670
- 69 Shimodaira, H. and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116
- 70 Huelsenbeck, J.P. *et al.* (1996) A likelihood-ratio test of monophyly. *Syst. Biol.* 45, 546–558
- 71 Goldman, N. (1998) Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. London Ser. B* 265, 1779–1786
- 72 Graybeal, A. (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47, 9–17
- 73 Yang, Z. (1998) On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* 47, 125–133
- 74 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556
- 75 Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (Vol. 3) (Munro, H.N., ed.), pp. 21–132, Academic Press