THE ROYAL
SOCIETY

Phylogenetic information and experimental design in molecular systematics

Nick Goldman

Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK (n.goldman@gen.cam.ac.uk)

Despite the widespread perception that evolutionary inference from molecular sequences is a statistical problem, there has been very little attention paid to questions of experimental design. Previous consideration of this topic has led to little more than an empirical folklore regarding the choice of suitable genes for analysis, and to dispute over the best choice of taxa for inclusion in data sets. I introduce what I believe are new methods that permit the quantification of phylogenetic information in a sequence alignment. The methods use likelihood calculations based on Markov-process models of nucleotide substitution allied with phylogenetic trees, and allow a general approach to optimal experimental design. Two examples are given, illustrating realistic problems in experimental design in molecular phylogenetics and suggesting more general conclusions about the choice of genomic regions, sequence lengths and taxa for evolutionary studies.

Keywords: experimental design; Fisher information; likelihood; molecular evolution; phylogenetic information; phylogenetics

1. INTRODUCTION

Phylogenetic inference from molecular sequences is increasingly being perceived as a statistical problem. Probabilistic models of nucleotide substitution or amino-acid replacement can be allied with maximum-likelihood (ML) inference to take advantage of established statistical theory (Goldman 1990), to enable model-fitting (Yang *et al.* 1994; Goldman *et al.* 1998), to provide great flexibility in testing evolutionary hypotheses (Huelsenbeck & Rannala 1997), and simply to give excellent results in the inference of evolutionary relationships (Kuhner & Felsenstein 1994; Huelsenbeck 1995). Constraints imposed by computational complexity are becoming less restrictive as computers improve. Given the recognition of phylogenetic inference as being inherently statistical in nature, it is surprising that so little attention has been paid to experimental design.

A number of topics of relevance to experimental design in phylogenetics have been discussed previously. There is something of a folklore surrounding the choice of genes or other genomic regions for investigating particular evolutionary questions, but almost no published quantitative results. It is widely accepted that sequences that have undergone very little evolutionary change since their divergence from a common ancestor, through low substitution rates or short evolutionary times, will exhibit too few differences to contain useful evolutionary information. Equally, sequences that have undergone very large amounts of change (high rates or long times) become 'saturated' with changes and no evolutionary signal is detectable amid the noise. Consequently, a happy medium is expected at some intermediate level of sequence divergence, but the 'asymptotic' results (for extreme high and low levels of divergence) give no clue as to where this lies. Although there is considerable experience with particular genes and organisms (see Hillis *et al.* (1996, pp. 336–339) for an extensive list of studies), where comparisons among genes have been made they tend to be evaluated empirically by the congruence of the results obtained among themselves and with researchers' *a priori* expectations. In addition, a large laboratory effort is needed before even these qualitative conclusions can be reached.

The methods introduced in this paper can quantify the effects of varying levels of divergence. They confirm the belief that intermediate levels of sequence divergence are most useful, and are able to give estimates of optimal levels of divergence. The necessary analyses can be done before any data are collected. The only other method for assessing which genomic regions are likely to be most useful in phylogenetic questions is that of Yang (1998). This uses simulation to estimate probabilities of successful tree-topology inference. The approach may be extremely time-consuming for realistic problems.

The choice of taxa to include in phylogenetic studies has also rarely been discussed. Li *et al.* (1987) considered the effects of adding outgroup taxa, and Ritland & Clegg (1990) and Maddison *et al.* (1992) are agreed that if outgroups are to be added, they should not be too distantly related to the ingroup taxa (a conclusion that is confirmed in this paper). More recently, consideration of the estimation of large phylogenies has led to the contradictory advice that the number of sequences included in a study be reduced (Kim 1996) and increased (particularly to break long branches in trees (Hillis 1996)) in order to improve inferences (see also Hillis 1998). Although it is not clear that the latter strategy will always be successful (Zharkikh & Li 1993; Kim 1996), Strimmer & von Haeseler (1996) and Hillis (1996) have

Received 28 May 1998 *Accepted* 8 June 1998

demonstrated that the estimation of large phylogenies to high accuracy is quite possible. Both Graybeal (1998) and Yang (1998) have used simulations to estimate probabilities of successful tree inference under conditions of varying numbers of taxa. In this paper I show how to quantify the information content of a data set. It will become apparent that augmenting a data set can only increase the information content with respect to any particular parameter of interest (viewed in isolation), although adding new sequences simultaneously increases the total number of parameters (e.g. branch lengths) to be estimated.

The length of (aligned) sequences to be analysed is also of relevance in experimental design. Churchill *et al.* (1992) developed a method for assessing the number of sites required to test reliably whether the relationships among four taxa were best described by a star phylogeny or a fully resolved tree, but it is not clear that this method can be extended to other evolutionary questions. An empirical and heuristic approach developed by Martin *et al.* (1995) seeks to estimate the number of sites needed to estimate phylogenetic relationships correctly, but the approach requires a fully analysed data set before any conclusions can be drawn and it does not permit extrapolation of results from the analysis of one combination of model phylogeny and genomic region to any other such combination. In contrast, the methods I introduce incorporate sequence length in a straightforward manner, allowing immediate quantification of its effects on information content.

There have been numerous studies whose primary aim has been to compare the performance of different phylogenetic inference methods (e.g. Kuhner & Felsenstein 1994; Huelsenbeck 1995). Typically, success is measured by the proportion of the time a method can recover the correct evolutionary relationships from simulated data. Components of some such studies' results can be interpreted in relation to questions in experimental design. The development of these methods into a comprehensive experimental design strategy as initiated by Graybeal (1998) and Yang (1998) might be difficult, however, not least because of the long computation times that result from repeated simulation and analysis of realistic data sets.

A number of methods, together classed as permutation tests, have been developed to assess whether or not data sets contain hierarchical structure (e.g. Swofford *et al.* 1996*a*). These methods have remained controversial (e.g. Swofford *et al.* 1996*b*). In any case, my interest in this paper is not to test for the existence of any hierarchical structure but to measure what information there is expected to be in a data set, with a view to designing more efficient experiments.

In this paper, I develop a general approach to experimental design in molecular phylogenetics. It uses standard optimal experimental design methods, and allows simultaneous consideration of taxon, genomic region and sequence-length selection. Measures are derived from the Fisher information matrix (Edwards 1972; Atkinson & Donev 1992), which quantify information with respect to parameters (branch lengths or positions of internal nodes of trees) and which are themselves of interest or may represent regions of a phylogeny that are of particular interest.

Although tree topology can be the parameter of most interest to systematists, it is not clear how it can be treated as a standard parameter in phylogenetic inference (Yang *et al.* 1995). The simulation approach developed by Graybeal (1998) and Yang (1998) can assess the probability of estimating an entire topology correctly, or the proportion of a topology that will on average be correctly inferred. In this paper I have taken a different approach, concentrating on other parameters of interest (branch lengths) and treating topology as a fixed part of the model much as, say, the choice between linear, polynomial or nonlinear models is fixed in more traditional experimental designs. As with that analogy, other forms of analysis (not considered in this paper) are needed to select the best among candidate models.

Within the unifying methodology developed here, it is possible to find optimal phylogenetic experimental designs within the constraints imposed by real experimental considerations. In the following sections, the necessary theory is developed for the case of four-state Markov models of DNA nucleotide substitution first without, and later with, the assumption of a molecular clock. Two example applications are described, followed by a discussion of the results and generalizations that can be derived from them. It may be worth stating explicitly that although all examples given here assume the Jukes & Cantor (1969) model of nucleotide substitution, this is by no means necessary and is not assumed in the notation used. The methods are also equally applicable to phylogenetic inference from amino-acid sequences, although this has not yet been implemented.

2. METHODS

The information matrix I of an experiment to estimate the vector parameter θ by ML can be defined by

$$I_{ij} = -\left(\frac{\partial^2 \ln(L)}{\partial\theta_i \partial\theta_j}\right) = -\left(\frac{\partial^2 S}{\partial\theta_i \partial\theta_j}\right) \qquad (1)$$

(Edwards 1972; Stuart & Ord 1991), where L is the likelihood function and S is the support, equal to $\ln(L)$. In the simplest case, θ is one-dimensional and the information is a scalar quantity,

$$I(\theta) = -\left(\frac{\partial^2 S}{\partial\theta^2}\right). \qquad (2)$$

Informally, $I(\theta)$ is proportional to the precision of an (unbiased) estimator of θ. The second derivative of the log-likelihood function (equation (1)) is a measure of the sharpness of the peak in the likelihood function. A sharp peak (large information) makes the location of the maximum easy and indicates confident estimation, whereas a relatively shallow peak (low information) indicates less certainty.

For the purposes of experimental design, it is appropriate to concentrate on the Fisher or expected information $E_{\theta^*}(I)$, which depends on θ^*, the true value of θ (Edwards 1972; Atkinson & Donev 1992). The expected information matrix is defined by

$$E_{\theta^*}(I_{ij}) = -E_{\theta^*}\left(\frac{\partial^2 S}{\partial\theta_i \partial\theta_j}\right). \qquad (3)$$

To simplify this notation, θ^* is omitted when there is no ambiguity. The inverse of the expected information matrix gives asymptotic lower bounds for the variances of estimators of the θ_i (Stuart & Ord 1991), and $E(I)$ is widely used in experimental design problems. A design criterion is defined in terms of the elements of $E(I)$, and different experimental designs for the same estimation problem can be assessed in terms of this measure. Different optimality criteria have been proposed, concentrating on different aspects of estimation problems. I shall consider two such criteria. The first considers the information relating to just one parameter in particular, irrespective of other parameters. This information is measured by the diagonal element of the expected information matrix corresponding to the chosen parameter, i.e. $E(I_{ii})$ if the chosen parameter is θ_i (cf. equation (2)).

The second criterion illustrated here is the determinant of $E(I)$, written $|E(I)|$, which is a measure of information pertaining to all parameters simultaneously. Relatively large values of this determinant indicate relatively large total amounts of information. Use of this measure for experimental design is known as the D-optimum criterion (Atkinson & Donev 1992). Its reciprocal, $|E(I)|^{-1}$, is equivalent to 'generalized variance' (Stuart & Ord 1991; Atkinson & Donev 1992).

In contrast, the observed information, $I_{ij} = -(\partial^2 S/\partial\theta_i\partial\theta_j)_{\hat{\theta}}$, depends on the observed data and on $\hat{\theta}$, the ML estimate of θ. The observed information measures the precision actually attained in an experiment, and its consideration is closely related to the use of the curvature method for estimating asymptotic confidence intervals for parameter estimates (Stuart & Ord 1991; Yang *et al.* 1995). It is not used for experimental design purposes.

In the context of phylogenetic inference, it is most convenient to develop theory in terms of the expected information per sequence site. As with all experimental design questions, the true parameter values are not known and, in order to proceed, a plausible experimental design phylogeny H must be assumed. We must consider all data 'patterns' that could appear at sites of an alignment of DNA sequences. Each pattern is one of the 4^n possible combinations of nucleotides A, C, G and T observable at a site for each of n sequences. I use $b = 1, 2, \ldots, 4^n$ to label these patterns, and p_b to denote their probabilities of occurrence given the experimental design phylogeny H. Initially, I develop the theory for the case of ML phylogenetic inference with no assumption of a molecular clock. In the absence of a molecular clock, the branch lengths of permitted phylogenetic trees can be varied independently.

For simplicity, all calculations reported in this paper use the Jukes–Cantor model of DNA substitution (Jukes & Cantor 1969), which has no free parameters. Consequently, the set of branch lengths of H form the parameters θ_i. The derivations below, however, do not require the Jukes–Cantor model and apply to any time-reversible Markov-process model of DNA substitution, represented by Q, its matrix of instantaneous rates (Swofford *et al.* 1996*a*). (If Q contained any of the parameters θ_i, e.g. a transition/transversion rate ratio parameter, then equations (11) and (13) below would be altered owing to the more complicated nature of derivatives with respect to these parameters.) There seems no reason to expect that results using other Markov-process models for DNA substitution will give qualitatively different results. The case of inference with the assumption of a molecular clock is developed below. It is not clear how the derivatives in the above equations can be extended to the (discrete) tree-topology parameter, and this has not been studied.

Writing $S_b = \ln(p_b)$, we obtain the following for the elements of $I(b)$, the information matrix for pattern b:

$$I(b)_{ij} = -\frac{\partial^2 S_b}{\partial\theta_i\partial\theta_j} = -\frac{1}{p_b}\frac{\partial^2 p_b}{\partial\theta_i\partial\theta_j} + \frac{1}{p_b^2}\frac{\partial p_b}{\partial\theta_i}\frac{\partial p_b}{\partial\theta_j}. \tag{4}$$

From equations (3) and (4), the expected information per site is given by

$$\begin{aligned} E(I_{ij}) &= -E\left(\frac{\partial^2 S}{\partial\theta_i\partial\theta_j}\right) = \sum_{b=1}^{4^n} p_b I(b)_{ij} \\ &= -\sum_{b=1}^{4^n}\frac{\partial^2 p_b}{\partial\theta_i\partial\theta_j} + \sum_{b=1}^{4^n}\frac{1}{p_b}\frac{\partial p_b}{\partial\theta_i}\frac{\partial p_b}{\partial\theta_j}. \end{aligned} \tag{5}$$

The p_b are constrained by $\sum p_b = 1$ and thus $\sum \partial^2 p_b/\partial\theta_i\partial\theta_j = 0$. Therefore,

$$E(I_{ij}) = \sum_{b=1}^{4^n}\frac{1}{p_b}\frac{\partial p_b}{\partial\theta_i}\frac{\partial p_b}{\partial\theta_j} \tag{6}$$

(Edwards 1972). The expected total information for sequences of N sites is N times this:

$$E^{\text{tot}}(I_{ij}) = NE(I_{ij}) = N\sum_{b=1}^{4^n}\frac{1}{p_b}\frac{\partial p_b}{\partial\theta_i}\frac{\partial p_b}{\partial\theta_j}. \tag{7}$$

The p_b and their derivatives can be calculated as follows. For a given (model) tree topology and branch lengths, define the set of all nodes to be $A = \{1, 2, \ldots, m, m+1, \ldots, m+n\}$, with nodes $1, 2, \ldots, m$ being the internal nodes of the tree and $m+1, m+2, \ldots, m+n$ being the tips. (For a fully bifurcating unrooted tree, $m = n-2$.) The ordering of nodes and tips within their respective subsets is arbitrary. Without loss of generality, we can define internal node $1 \in A$ to be the 'root' of the tree. This does not affect the probability calculations for any reversible Markov-process model of nucleotide substitution (Felsenstein 1981), and defines a direction (away from the root) on each branch of the tree. The branches can now be uniquely labelled $x \in B = \{2, 3, \ldots, m+n\}$ according to the node or tip they lead to, and their lengths are then denoted θ_x. The topology and root define a relationship on A whereby we write $x \rightsquigarrow y$ $(x, y \in A)$ if there is a branch directly joining x and y with the direction from x to y being away from the root, and write $x \not\rightsquigarrow y$ otherwise. Writing $b_x \in \{\text{A, C, G, T}\}$ for the nucleotide (observed or unobserved) at $x \in A$, we define

$$q_{b_x b_y}(\theta_y) = \begin{cases} (\mathrm{e}^{\theta_y Q})_{b_x b_y} & : \quad x \rightsquigarrow y \\ 1 & : \quad x \not\rightsquigarrow y. \end{cases} \tag{8}$$

For pairs $x, y \in A$ such that $x \rightsquigarrow y$, this equals the probability that a branch of length θ_y with nucleotide b_x at one end has nucleotide b_y at the other end (Bartlett 1978; Swofford *et al.* 1996*a*).

Given these definitions,

$$p_b = \left(\prod_{k=1}^{m}\sum_{b_k}\right)\left(\pi_{b_1}\prod_{x,y\in A} q_{b_x b_y}(\theta_y)\right), \tag{9}$$

where we have used the notation

$$\left(\prod_{k=1}^{m}\sum_{b_k}\right) = \left(\sum_{b_1}\sum_{b_2}\cdots\sum_{b_m}\right), \tag{10}$$

with all summations being over all nucleotides, $b_k \in \{\text{A, C, G, T}\}$.

For each branch $j \in B$, there is a unique node i such that $i \rightsquigarrow j$ and then

$$\frac{\partial p_b}{\partial \theta_j} = \left(\prod_{k=1}^{m}\sum_{b_k}\right)\left(\left(\pi_{b_1}\prod_{\substack{x,y\in A\\(x,y)\neq(i,j)}} q_{b_x b_y}(\theta_y)\right)\frac{\partial q_{b_i b_j}(\theta_j)}{\partial \theta_j}\right), \tag{11}$$

because θ_j appears in exactly once in equation (9), in the term $q_{b_i b_j}(\theta_j) = (\mathrm{e}^{\theta_j Q})_{b_i b_j}$. Standard properties of Markov chains (Bartlett 1978) give

$$\frac{\partial q_{b_i b_j}(\theta_j)}{\partial \theta_j} = \left(Q e^{\theta_j Q}\right)_{b_i b_j}, \tag{12}$$

and thus

$$\frac{\partial p_b}{\partial \theta_j} = \left(\prod_{k=1}^{m}\sum_{b_k}\right)\left(\left(\pi_{b_1}\prod_{\substack{x,y\in A\\(x,y)\neq(i,j)}} q_{b_x b_y}(\theta_y)\right)(Q\mathrm{e}^{\theta_j Q})_{b_i b_j}\right). \tag{13}$$

Felsenstein's (1981) 'pruning algorithm' is an efficient way of performing the calculation of the p_b given by equation (9). The similarity of equations (9) and (13) indicate ways in which the derivatives can be calculated much as the probabilities are, and many partial calculations from the probabilities can be re-used to give the derivatives at little extra computational cost.

In the case of analysis with the assumption of a molecular clock, the tree has a natural root node and there are the restrictions that the distances from this root to each of the tips must be equal (Felsenstein 1981). These constraints mean that branch lengths cannot be varied independently of one another, and are no longer appropriate parameters in information calculations. It is useful instead to consider the positions of internal nodes of a tree relative to its tips as the parameters of interest. This alternative parameterization is convenient, as independent movement of internal nodes (including the root) nearer to or further from their ancestors is permitted.

Fortunately, little additional computation is needed to perform this change in parameters. If the calculation of $E(I(\theta))$ is first completed as above without the constraints associated with a clock-like tree, re-parameterization can proceed as follows. Consider, for example, the position illustrated in figure 1. The original (unconstrained) parameters $(\theta_1, \theta_2, \theta_3)$ cannot vary independently of one another when the tree is interpreted as clock-like (θ_1 and θ_2 must be equal if the root is in the branch with length θ_3). Two new parameters (ψ_1, ψ_2), representing the positions of the two internal nodes of the clock-like tree relative to the tips of the tree, can be defined in terms of their directions in $(\theta_1, \theta_2, \theta_3)$-space. A unit increase in ψ_1 (moving node 1 away from the tips) corresponds to unit increases in θ_1 and θ_2, and a simultaneous unit decrease in θ_3. In vector notation, ψ_1 is represented by the vector $T_1 = (1, 1, -1)$ in $(\theta_1, \theta_2, \theta_3)$-space. Similarly, ψ_2 is represented by $T_2 = (0, 0, 2)$ (moving node 2 away from the tips induces no change in θ_1 or θ_2, but requires θ_3 to be increased on both sides of the root node 2).

In general, we can write T_i to indicate the direction in θ-space of each new parameter ψ_i and T for the matrix with ith row T_i. Standard vector calculus results on directional derivatives (Spiegel 1959) then give

$$\frac{\partial^2 S_b}{\partial \psi_i \partial \psi_j} = \sum_{k,l} T_{ik}\frac{\partial^2 S_b}{\partial \theta_k \partial \theta_l} T'_{lj}\,, \tag{14}$$

where the prime symbol (′) indicates matrix transposition. Hence, from equation (3),

$$E(I(\psi)) = TE(I(\theta))T'. \tag{15}$$

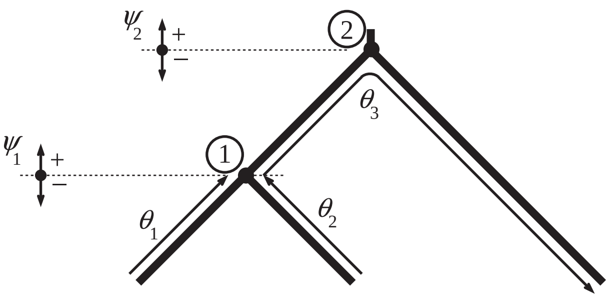

Figure 1. Transformation of variables for the case of rooted trees. The three-species tree with branch length parameters $\theta = (\theta_1, \theta_2, \theta_3)$ is clock-like when θ_1 is constrained to equal θ_2. In this case, the positions of the internal nodes of the rooted tree (circled 1 and 2) form the new parameters $\psi = (\psi_1, \psi_2)$.

Branch lengths θ_i are the products of a rate μ and times t_i. The t_i and μ are confounded and cannot normally be estimated (or known) separately (Felsenstein 1981; Swofford *et al.* 1996*a*). This can be of importance when using equations (6) and (7). If the experimental design question being considered does not involve any variation in evolutionary rate, μ is effectively an (unknown) constant. In this case we may use equations (6) and (7) as above, as information regarding parameters θ_i is equivalent to information regarding t_i. It is possible, however, that experimental designs under consideration involve varying values of μ while assessing information regarding the t_i (for example, in relation to calibrating phylogenies to estimate divergence dates). In this case, derivatives with respect to θ_i in equations (4–7) should be replaced by derivatives with respect to t_i. Because $t_i = \theta_i/\mu$, we immediately obtain

$$E(I_{ij}) = \mu^2 \sum_{b=1}^{4^n} \frac{1}{p_b}\frac{\partial p_b}{\partial \theta_i}\frac{\partial p_b}{\partial \theta_j} \tag{16}$$

and

$$E^{\mathrm{tot}}(I_{ij}) = \mu^2 N E(I_{ij}) = \mu^2 N \sum_{b=1}^{4^n} \frac{1}{p_b}\frac{\partial p_b}{\partial \theta_i}\frac{\partial p_b}{\partial \theta_j}. \tag{17}$$

In other words, if the experimental designs under consideration involve variation in the relative rate of evolution μ, then the original information matrix elements should be multiplied by μ^2 to give values comparable across different values of μ. In the case of a molecular clock, equation (15) remains valid after the elements of $E(I(\theta))$ are altered appropriately.

3. EXAMPLES

(a) *Best experimental design in adding sequence to an existing alignment (information relating to one node of a clock-like phylogeny)*

Figure 2*a* depicts a rooted phylogeny for five sequences. This is modelled on the phylogeny obtained by applying the Jukes–Cantor model of DNA substitution, with the assumption of a molecular clock, to the aligned 895-base pair (bp) mtDNA sequences of Brown *et al.* (1982), as studied by Bishop & Friday (1985) and Yang *et al.* (1995). This tree is taken as the model, and we imagine the case that we are interested in increasing the total information $E^{\mathrm{tot}}(I_G)$ relating to the position of the node G at which the gorilla lineage diverges. We consider two options:

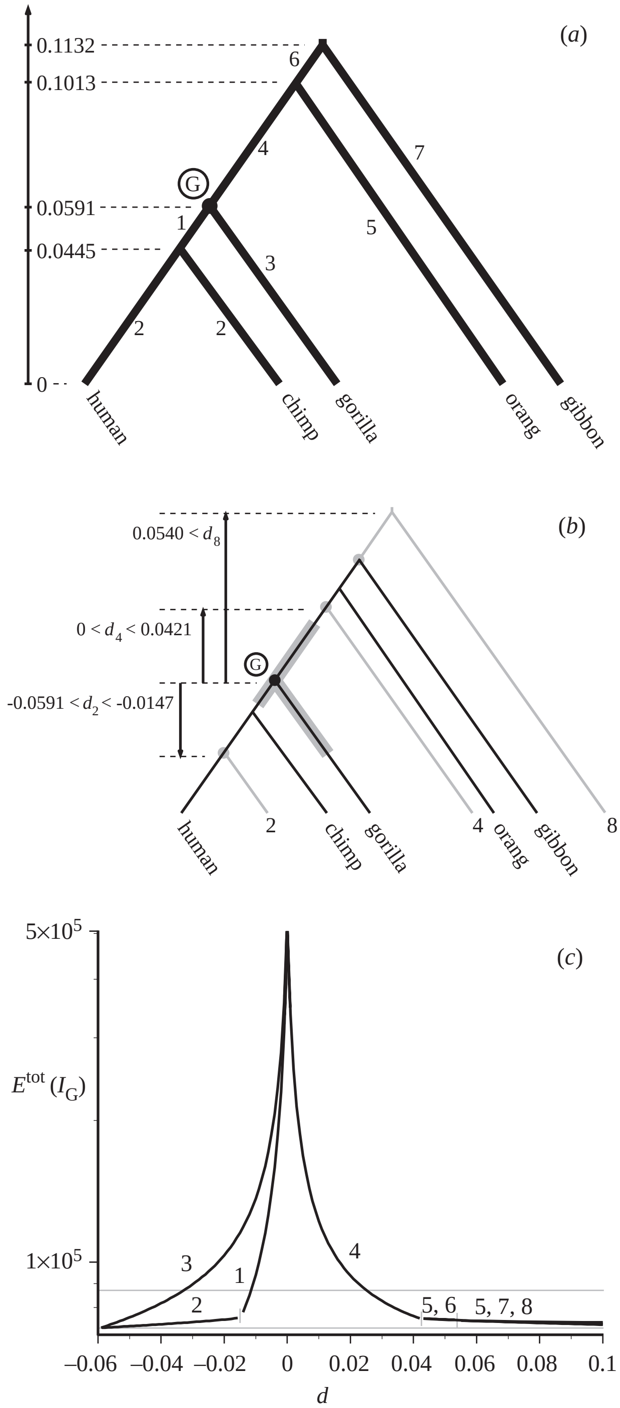

Figure 2. Adding sequence to an existing alignment. (*a*) The model phylogeny, indicating the node of particular interest (circled *G*) as described in the text and with branches labelled 1–7. (The two branches labelled 2 are equivalent in this example.) (*b*) The same phylogeny with three examples of possible additional sequences (greyed branches and labels). These are cases 2, 4 and 8 as described in the text. Corresponding permitted values of d_2, d_4 and d_8 are also indicated. A new sequence joining the model tree within the greyed area around node *G* increases the information regarding that node more than increasing the lengths of the original sequences by an equivalent amount would. (*c*) The information measure $E^{\text{tot}}(I_G)$ plotted against *d*, the distance between the node *G* and the point at which a new sequence attaches to the model tree. Regions of the graph are labelled 1–7 according to which branch of the model tree the new sequence is attached to, or 8

increasing the lengths of each of the current five sequences by 179 bp, or adding a sequence from an extant species. Either choice requires the sequencing of an additional 895 bp; we assume the 'cost', in terms of laboratory effort, is equal for these two strategies.

Because the design options considered do not involve variation of μ, we apply equations (7) (equivalently, equation (17) with constant $\mu = 1$) and (15). The 'baseline' total information relating to node *G*, before any additional sequence is added, is $895 \times 80.89 = 7.240 \times 10^4$. If the sequences were each increased by 179 bp to 1074 bp, this would become $1074 \times 80.89 = 8.687 \times 10^4$. This is the value that must be exceeded by the addition of a new 895-bp sequence if this latter strategy is to be more beneficial in increasing the information about the node *G* at a fixed cost.

If an additional sequence is to be added to the phylogeny of figure 2*a*, it will either be on one of the branches labelled 1–7 or (case 8) be an outgroup to all five sequences already present. By symmetry, it is clear that the two branches sharing the label 2 are equivalent in this respect and need not be considered individually. The cases 1–8 may be placed on a linear scale *d* that measures the distance between node *G* and the position at which the new sequence joins the model tree. The value of *d* is defined to be negative for cases 1–3 (joining 'below' *G*) and positive for cases 4–8 (joining 'above' *G*). For an outgroup sequence, case 8, it is appropriate to measure the distance from *G* to the new root of the tree. Thus, in this example *d* can take values between -0.0592 (at the descendent ends of branches 2 and 3) and $+\infty$ (case 8, for an exceedingly ancient outgroup). Where ambiguity is possible, *d* is given a subscript indicating the branch to which it refers. Three possible positions for a new sequence (cases 2, 4 and 8), and indications of their corresponding values of *d* (d_2, d_4 and d_8) are shown in figure 2*b*.

Calculations of $E^{\text{tot}}(I_G)$ were made for $-0.0592 \leqslant d \leqslant 0.1$. These are indicated in figure 2*c*. Also indicated in this figure are the baseline values attained with no extra sequence, and by simply increasing the length of each of the original five sequences by 179 bp. As would be expected, if a sequence is added that is a duplicate of one of the original five ($d = -0.0592$; also $d_5 = 0.1434$ and $d_7 = 0.1672$, results not shown) then there is no increase in information. The nearer to *G* that the additional sequence joins the tree, the greater is the gain in information, rising to a maximum if the new sequence joins the model tree exactly at *G* ($d = 0$).

In particular, notice that the baseline value available by increasing the length of the original sequences is exceeded for $-0.0115 < d_1 \leqslant 0$, for $-0.0330 < d_3 \leqslant 0$, and for $0 \leqslant d < 0.0263$. These three regions (indicated by the grey-shaded areas on the tree of figure 2*b*) contain those points within the model tree such that an additional (contemporary) 895-bp sequence joining there would

(*Continued*) if the new sequence is an outgroup. Also indicated is the information content of the original sequences of 895 bp (lower grey line) and the information if these sequences are each increased to 1074 bp (upper grey line). A logarithmic scale is used for $E^{\text{tot}}(I_G)$ to improve clarity.

give more additional information about node G than would augmentation of the original sequences by a total of 895 bp. In this particular example, because these regions indicate either a species placed phylogenetically between the orang-utan and the divergence of human and chimp, or one that is a sister group to gorilla (but not too closely related, as for example a second gorilla sequence would be), it seems unlikely that such a specimen could be found and laboratory resources would be better used extending the original five sequences. For example, note that any additional outgroup sequence (case 8, $d_8 \geqslant 0.0541$) adds virtually no information about node G.

(b) *Best experimental design in choice of ideal gene for sequencing (information relating to all branch lengths of an unrooted phylogeny)*

Figure 3*a* depicts an unrooted phylogeny modelled on that obtained by analysing 6166 bp $\psi\eta$-pseudogene sequences of human, chimp, gorilla, orang-utan, rhesus monkey and spider monkey (Fitch *et al.* 1988; Yang *et al.* 1995) using the Jukes–Cantor model of DNA substitution. In this example, we consider the question of what gene, or more specifically what nucleotide substitution rate (μ) for a gene, might give more information per site regarding the branch lengths (t_i) of this model phylogeny. This is done by scaling the entire tree (i.e. all branch lengths) by some factor μ, to represent a gene with evolutionary rate μ times that of the $\psi\eta$-pseudogene. Scaling the tree by varying factors allows us to investigate the amount of information per site relating to all branch lengths, $|E(I)|$, as a function of the rate of substitution relative to that of the $\psi\eta$-pseudogene sequence.

In this case, we apply equation (16). In practice, it is convenient to look at $|E(I)|^{1/9}$, which can be interpreted as a measure of the average amount of information per branch-length parameter for this nine-branch tree. Values of $|E(I)|^{1/9}$ for μ in the range 0.1–100 are plotted in figure 3*b*. We find that $|E(I)|$ has a maximum at $\mu = 10.6$. In this example, therefore, the 'ideal' gene for phylogenetic study (as indicated by the $|E(I)|$ information measure) would be one with rate $\mu = 10.6$ times that of the $\psi\eta$-pseudogene.

Of course, it will not always be possible to select a gene with precisely the required rate of evolution. Graphs such as figure 3*b* can then be of assistance in selecting the best available gene. In this example, it is interesting to note that genes evolving at a much greater rate would still contain relatively large amounts of information (e.g. $\mu = 37$ gives approximately the same $|E(I)|$ as does $\mu = 1$). Any gene with $1 < \mu < 37$ (relative to the $\psi\eta$-pseudogene) would be more informative than the $\psi\eta$-pseudogene.

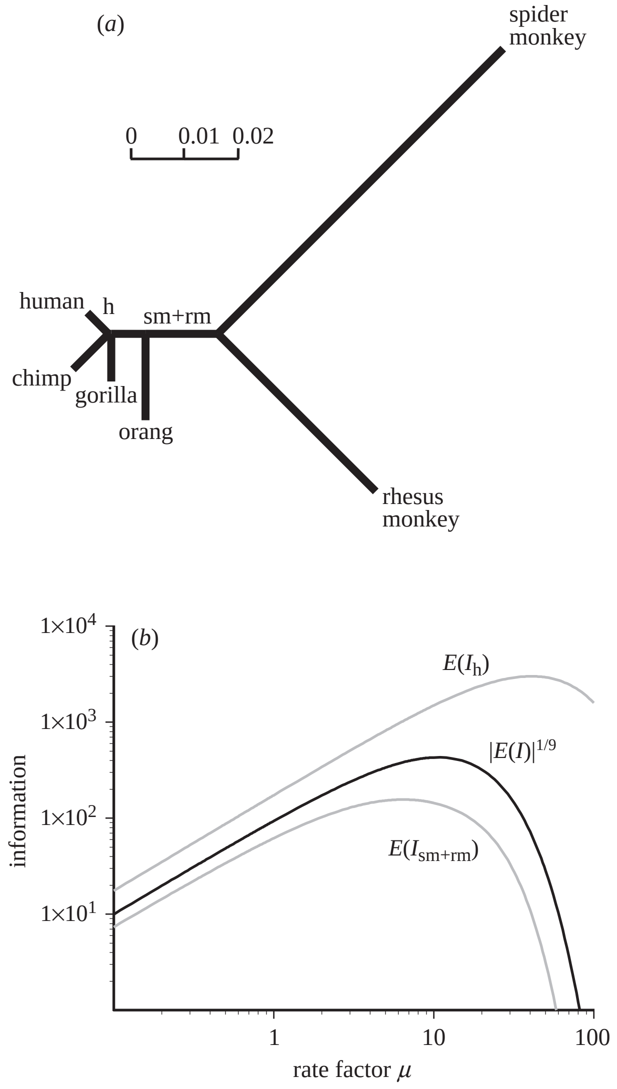

Figure 3. Choice of ideal gene for sequencing. (*a*) The model phylogeny. Branches referred to in the text as h and sm+rm are indicated. (*b*) Information plotted against μ, the rate of nucleotide substitution relative to that of the model phylogeny. The middle curve is the expected information $|E(I)|^{1/9}$. The upper and lower curves (grey) show $E(I_{\mathrm{h}})$ and $E(I_{\mathrm{sm+rm}})$, the information relating to branches h and sm+rm, respectively. A logarithmic scale is used for μ to improve clarity.

4. DISCUSSION

The first example in § 3(a) considers an experimental design question regarding the addition of sequence data. Figure 2*c* indicates a number of interesting features of this type of problem. The fact that the information relating to a parameter does not fall below the baseline level whenever sequence is added indicates that additional data, however remotely related to the parameter of interest, will never make inferences about that parameter worse. This appears to refute the recommendation (Kim 1996) that data might be excluded to improve phylogenetic estimates; this recommendation perhaps resulted from placing undue emphasis on estimating phylogenies entirely correctly (Yang & Goldman 1997; Hillis 1998). It is perhaps intuitively obvious that additional sequences should ideally join the model phylogeny as near to the node of interest as possible. It might not be so obvious, without the quantification of information that these new methods permit, as to quite how close to node G a new sequence must join in order to provide more information than simply lengthening the original sequences by a corresponding amount would. In particular, the option of adding any outgroup sequence is evidently of very little use in this example.

The second example, § 3(b), illustrates another realistic question in phylogenetics. Graphs such as figure 3*b* have the expected shape of low information at the highest and lowest levels of sequence divergence and a maximum at some intermediate level, and may be useful in the design of phylogenetic studies. Figure 3*b* also shows the expected information relating to the branches leading to human (h) and separating the spider and rhesus monkeys from the other species (sm+rm), $E(I_{\mathrm{h}})$ and $E(I_{\mathrm{sm+rm}})$, respectively. These curves have their maxima at $E(I_{\mathrm{h}}) = 40.8$ and $E(I_{\mathrm{sm+rm}}) = 6.45$, according with intuition and common experience that recent events (branch h) are best studied via relatively fast-evolving sequences and more ancient events (branch sm+rm) via more slowly evolving sequences. Note that results of this sort are not available using the approach of Graybeal (1998) and Yang (1998).

One aspect of phylogenetic inference that these new methods do not take into consideration is sequence alignment. It has recently been noted that alignment procedures can have significant effects on inferred phylogenies (Morrison & Ellis 1997; Goldman 1998). Experimental design studies might suggest, for example, the use of genes with levels of divergence sufficiently high that it would be difficult to align the sequences correctly. In this case the information measures described in this paper might overestimate the usefulness of the supposedly optimal sequences. Further work is needed to assess this possibility.

The assumptions I have used regarding 'laboratory costs' of obtaining sequence data are deliberately simple, for illustrative purposes (see also Pluzhnikov & Donnelly 1996). However, the method for assessing information content that I have developed here is flexible enough to allow more realism, and its probabilistic basis could allow even complex cost functions and decision-theoretic issues (Lindley 1985) to be taken into consideration.

I am grateful to Tim Massingham for writing computer programs implementing the methods described in this paper and for the original derivation of equation (13), while supported during the summer of 1997 by a Nuffield Foundation Undergraduate Research Bursary. A. W. F. Edwards provided enlightenment on Fisher information. Edward Holmes kindly provided assistance with, and Adrian Friday and Ziheng Yang made useful comments on, earlier versions of the paper. Z. Yang pointed out errors in the application of the method to the second example, and Matthew Stephens inspired their correction. N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research. Computer software (written in C) performing the analyses described in this paper is available on request.

REFERENCES

Atkinson, A. C. & Donev, A. N. 1992 *Optimum experimental designs.* Oxford University Press.

Bartlett, M. S. 1978 *An introduction to stochastic processes.* Cambridge University Press.

Bishop, M. J. & Friday, A. E. 1985 Evolutionary trees from nucleic acid and protein sequences. *Proc. R. Soc. Lond.* B **226**, 271–302.

Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Molec. Evol.* **18**, 225–239.

Churchill, G. A., von Haeseler, A. & Navidi, W. C. 1992 Sample size for a phylogenetic inference. *Molec. Biol. Evol.* **9**, 753–769.

Edwards, A. W. F. 1972 *Likelihood.* Cambridge University Press.

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evol.* **17**, 368–376.

Fitch, D. H. A., Mainone, C., Slightom, J. L. & Goodman, M. 1988 The spider monkey $\psi\eta$-globin gene and surrounding sequences: recent or ancient insertions of LINEs or SINEs? *Genomics* **3**, 237–255.

Goldman, N. 1990 Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.* **39**, 345–361.

Goldman, N. 1998 Effects of sequence alignment procedures on estimates of phylogeny. *Bioessays* **20**, 287–290.

Goldman, N., Thorne, J. L. & Jones, D. T. 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.

Graybeal, A. 1998 Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17.

Hillis, D. M. 1996 Inferring complex phylogenies. *Nature* **383**, 130–131.

Hillis, D. M. 1998 Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* **47**, 3–8.

Hillis, D. M., Mable, B. K., Larson, A., Davis, S. K. & Zimmer, E. A. 1996 Nucleic acids IV: sequencing and cloning. In *Molecular systematics*, 2nd edn (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 321–381. Sunderland, MA: Sinauer.

Huelsenbeck, J. P. 1995 Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48.

Huelsenbeck, J. P. & Rannala, B. 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227–232.

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism*, vol. 3 (ed. H. N. Munro), pp. 21–132. New York: Academic Press.

Kim, J. 1996 General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* **45**, 363–374.

Kuhner, M. K. & Felsenstein, J. 1994 A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molec. Biol. Evol.* **11**, 459–468. (See also Erratum. *Molec. Biol. Evol.* **12**, 525 (1994).)

Li, W.-H., Wolfe, K. H., Sourdis, J. & Sharp, P. M. 1987 Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 847–856.

Lindley, D. V. 1985 *Making decisions*, 2nd edn. London: John Wiley.

Maddison, D. R., Ruvolo, M. & Swofford, D. L. 1992 Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst. Biol.* **41**, 111–124.

Martin, M. J., Gonza´lez-Candelas, F., Sobrino, F. & Dopazo, J. 1995 A method for determining the position and size of optimal sequence regions for phylogenetic analysis. *J. Molec. Evol.* **41**, 1128–1138.

Morrison, D. A. & Ellis, J. T. 1997 Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Molec. Biol. Evol.* **14**, 428–441.

Pluzhnikov, A. & Donnelly, P. 1996 Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262.

Ritland, K. & Clegg, M. 1990 Optimal DNA sequence divergence for testing phylogenetic hypotheses. In *Molecular evolution. UCLA symposia on molecular and cellular biology, new series*, vol. 122 (ed. M. T. Clegg & S. J. O'Brien), pp. 289–296. New York: Wiley-Liss.

Spiegel, M. R. 1959 *Theory and problems of vector analysis and an introduction to tensor analysis.* New York: McGraw-Hill.

Strimmer, K. & von Haeseler, A. 1996 Accuracy of neighbor joining for *n*-taxon trees. *Syst. Biol.* **45**, 516–523.

Stuart, A. & Ord, J. K. 1991 *Kendall's advanced theory of statistics*, vol. 2, 5th edn. London: Edward Arnold.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996*a* Phylogenetic inference. In *Molecular systematics*, 2nd edn. (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 407–514. Sunderland, MA: Sinauer.

Swofford, D. L., Thorne, J. L., Felsenstein, J. & Wiegmann, B. M. 1996*b* The topology-dependent permutation test for monophyly does not test for monophyly. *Syst. Biol.* **45**, 575–579.

Yang, Z. 1998 On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* **47**, 125–133.

Yang, Z. & Goldman, N. 1997 Are big trees indeed easy? *Trends Ecol. Evol.* **12**, 357.

Yang, Z., Goldman, N. & Friday, A. 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molec. Biol. Evol.* **11**, 316–324.

Yang, Z., Goldman, N. & Friday, A. 1995 Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**, 384–399.

Zharkikh, A. & Li, W.-H. 1993 Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. *Syst. Biol.* **42**, 113–125.

As this paper exceeds the maximum length normally permitted, the author has agreed to contribute to production costs.