M. Feldman

Molecular Evolution, pages 289–296

OPTIMAL DNA SEQUENCE DIVERGENCE FOR TESTING PHYLOGENETIC HYPOTHESES

Kermit Ritland and Michael Clegg[1]

Department of Botany, University of Toronto
Toronto, Ontario M5S1A1 Canada
and
Department of Botany and Plant Sciences
University of California, Riverside, California 92521

ABSTRACT The information properties of the maximum likelihood phylogeny of a small number of DNA sequences is studied by inversion of Fisher information matrices. In a three–sequence phylogeny, an "optimal" sequence divergence, which minimizes the estimation variance of branch length ratios, effectively occurs over a broad range, ca. 25% to 50% divergence (of a maximum of 75%) between pairs of sequences. Increasing the divergence of the third sequence (perhaps an "outgroup") increases variance of estimated lengths and ratios of the first two branches. In a four–sequence tree, the optimum branch length decreases and occupies a narrower range, and variances are greater. These results are relevant for studies of relative rates of evolution in phylogenies.

INTRODUCTION

Comparative analyses of DNA sequence data provide inferences about both the phylogeny of genes and the pattern of substitutional changes within regions of a gene. For most purposes, maximum likelihood estimation (MLE) is the method of choice for statistical inferences. The likelihood method efficiently uses all information in the data, it rests upon a probabilistic model of mutational change, and its asymptotic statistical properties are quite easily studied regardless of the probability model used.

---

[1]This work was supported in part by a Natural Sciences and Engineering Research Council of Canada grant to KR, and in part by National Science Foundation grant BSR–8500206 to MC.

Given the probability model, the likelihood method seeks parameter values that maximize the probability of observed data. The likelihood function over the range of parameter values defines a likelihood surface, whose global maximum corresponds to the MLE parameter estimates. An attractive feature is that the variance of maximum likelihood estimates is related to the curvature of the average likelihood surface at the maximum (1), at least for large samples. In this paper, these variances will allow us to determine "optimal" levels of DNA sequence divergence to test evolutionary hypotheses.

The notion of "optimal" divergence is useful because different genomic components evolve at different rates (2). For example, introns, exons and flanking regions of genes, as well as different regions of the ribosomal RNA coding unit, are observed to evolve at different rates owing to different patterns of functional constraint. In addition, organelle genomes such as vertebrate mitochondrial DNA and plant chloroplast DNA evolve at markedly different rates relative to nuclear sequences (3,4). If the power to make evolutionary inferences depends upon sequence divergence, investigators may benefit by choosing among slowly vs. rapidly evolving sequences to address problems in evolution and systematics.

In this article, we use the known statistical properties of likelihood to investigate a series of questions about the testing of phylogenetic hypotheses. Although only simple trees are investigated, we note that many issues in DNA sequence analysis, for example the Human–Chimpanzee–Gorilla divergence, are based upon simple trees, and many principles derived from smaller trees are apt to hold for larger trees. Yet, the statistical properties of likelihood trees, whether simple or complex, are rarely studied (2,5).

A PROBABILITY MODEL FOR MUTATIONAL CHANGE

When applied to DNA sequences, the likelihood method requires a probability model of base substitution. For the present, we adopt the simplest model which assumes that (a) each nucleotide site has a constant probability of mutation, (b) different sites mutate independently and (c) replacement bases are randomly drawn from a pool with frequencies equal to those in the sequence (note that mutation may preserve the same nucleotide at a site). Following Felsenstein (6), if the base frequencies of A,T,C and G are π_i, i=1,2,3,4, then in a two-sequence comparison, the probability of observing base j in the second sequence, given base i in the first sequence, is

$$P_{ij}(t\mu,\pi) = \delta_{ij}\exp(-t\mu) + \pi_j(1-\exp(-t\mu)) \ ,$$

where t is the time separating the two sequences, μ is the mutation rate, and δ_{ij} is an indicator variable (δ_{ij}=1 if i=j and δ_{ij}=0 if i≠j). The number of substitutions expected at a site is $t\mu$. For three sequences, the probability of observing bases i, j, and k is the sum over possible "node" bases n,

$$\sum_n \pi_n \ P_{ni}(t_1\mu_1,\pi) \ P_{nj}(t_2\mu_2,\pi) \ P_{nk}(t_3\mu_3,\pi) \ ,$$

where this "node" is the ancestral sequence from which these sequences diverge (this is the trifurcation in an unrooted, three-sequence phylogeny; see diagram next page). The t's are the times that separate the three contemporary sequences from this node, so the $t_i\mu_i$ are the "branch lengths" in a phylogeny. More sequences involve analogous formula, with additional, nested summations over additional nodes (6,7).

MODUS OPERANDI

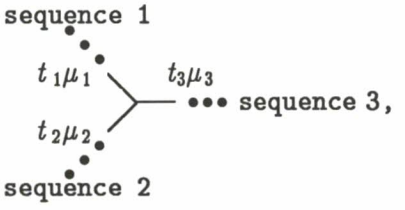
The approach we take is to compute variances, via inversion of the information matrix, over a range of parameter values, and to locate parameter values that give lowest variances. The information matrix contains expected second derivatives of the log-likelihood function (1), and in its computation, we assumed equality of base frequencies π_i. Derivatives were taken with respect to the products $t_i\mu_i$, i=1..3, and variances of other quantities were subsequently found via differential approximations. Variances are given on a per-site basis; if the sequence is n nucleotides long, variances for the sequence are obtained by dividing by n.

In the following, we distinguish between "sequence divergence": the observed proportion of sites that differ between two sequences, and "evolutionary divergence": the average number of substitutions at a site over evolutionary time (these include multiple substitutions).

It is obvious, at least to the statistically minded, that the variance, or uncertainty, of our estimate of evolutionary divergence of two sequences is a steadily increasing function of their sequence divergence. The goal of this paper was to examine another dimension: the uncertainty about ratios of evolutionary divergence. Such ratios correspond to the relative branch lengths, or "shape", of a phylogeny.

STATISTICAL PROPERTIES AND PRINCIPLES

The problem to be investigated concern the value of sequence divergence at which the variances are a minimum for a three sequence, unrooted topology. We first examine an unrooted topology over increasing values of $t_i\mu_i$,



assuming the evolutionary distances, $t_i\mu_i$, are equal.

Figure 1 shows the variance of estimates for three cases which we discuss. First, this figure shows the variance of absolute branch length (evolutionary distance $t_i\mu_i$) increases with sequence divergence, and it increases rapidly at higher levels of divergence.

The quantities of primary interest to us are variances of estimates of "relative" branch length and of branch length ratios. The former is $\langle t_i\mu_i\rangle/(\langle t_1\mu_1\rangle+\langle t_2\mu_2\rangle+\langle t_3\mu_3\rangle)$, i=1,3 (brackets denote estimates), and its estimation variance is

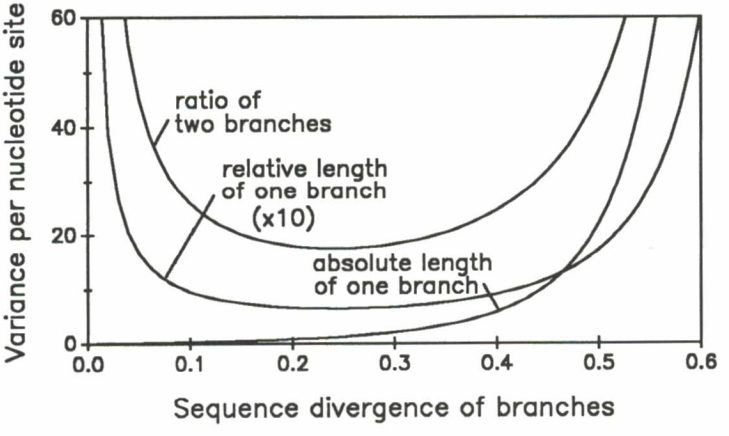


FIGURE 1. Variance, per nucleotide site, of estimates for a three-sequence tree, in relation to increasing sequence divergence between tips and node (maximum 0.75 divergence).

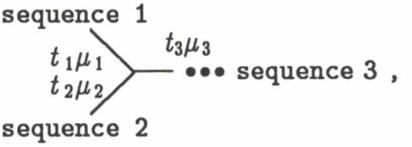
plotted against increasing sequence divergence in Figure 1 (variance is the same for all three branches). A minimum variance occurs at 25% divergence (75% is the maximum possible divergence), but note this minimum is shallow and effectively spans a wide range, from 10% to 40% divergence. A 25% divergence between tip and node corresponds to 42% pairwise divergence between two sequences (two tips) when multiple hits are taken into account. This minimum of variance, or "optimal" sequence divergence, occurs because as sequences diverge, the proportion of highly informative "single–hit" changes reaches a maximum relative to the less informative class of "multiple–hits" at a site.

The second quantity of primary interest is the variance in the estimate of the ratio of two branches. Frequently, relative rate tests are used to test for equality of rates of evolution in a topology. It is of interest to determine the sequence divergence for which this variance is minimized, as this is where relative rate tests would have the most statistical power.

If the first two sequences are more closely related to each other than to the third sequence, then $t_1=t_2$, and a test for differences in mutational rate between branches 1 and 2 is whether the ratio of estimated branch lengths, $\langle t_1\mu_1\rangle/\langle t_2\mu_2\rangle = \langle\mu_1\rangle/\langle\mu_2\rangle$, differs from one. The estimation variance for this ratio is also plotted Figure 1.

This ratio has an minimum variance near 25% divergence, and the "optimum" has a effective range from ca. 15% to 35% divergence from the node (this corresponds to a range of 27% to 54% pairwise sequence divergence). Thus, tests of molecular clocks are powerful over a wide range of sequence divergence. However, it is important to consider some of the biases that affect these calculations. Several violations of model assumptions will cause the minimum to be reduced over that shown in Figure 2, including (a) different rates of evolution for different regions of the nucleotide sequence, (b) different transition and transversion mutation rates, and (c) base frequencies that differ markedly from the 0.25 values assumed. In addition, problems of sequence alignment for highly diverged sequences can introduce another component of error. For all these reasons, it seems prudent to "choose" pairwise sequence divergence values near the lower end of the range, near 25–30%, when employing relative rate tests.

The second problem we consider focuses upon the effect of the distance of the third sequence upon the above variances. This sequence can be considered the "outgroup". We examine an unrooted three–sequence topology over increasing $t_3\mu_3$,



assuming the true branch lengths of $t_1\mu_1=t_2\mu_2=0.2$. Figure 2 shows that variances of relative individual branch lengths, and of ratios of branch lengths, increase monotonically with increased divergence of the outgroup. By contrast, the variance of the sum of two individual branch lengths remains constant. These results indicate that, for relative rate tests, outgroups should be closely related to the sequences of interest.

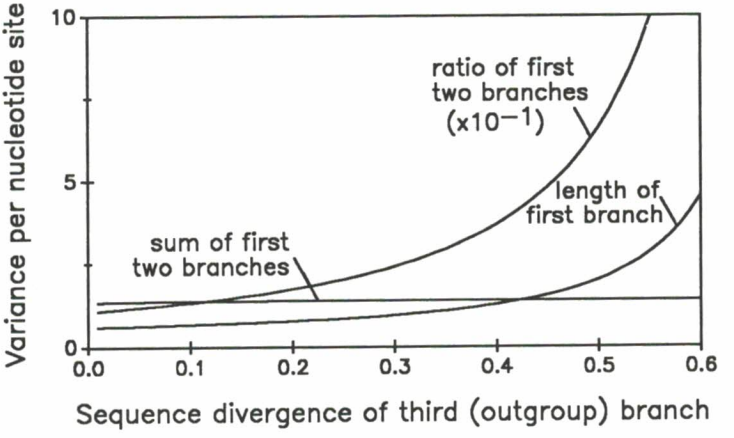


FIGURE 2. Variance of estimates for a three-sequence tree, in relation to greater distance of the outgroup branch.

The final problem involves a study of four-sequence trees. We address the same issues as in Figures 1-2, and see if our results hold for larger trees. First, we examine an unrooted four-sequence tree over increasing branch lengths, assuming equal branch lengths $t_i\mu_i$. This four-sequence tree, depicted in the following page, has two nodes, so there are two types of branches, "node-tip" branches (branches 1-4) and an "internode" branch (branch 5):

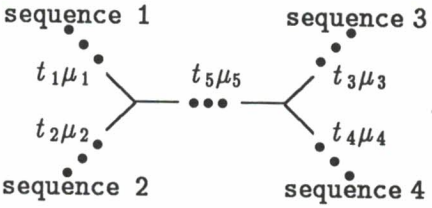


Figure 3 presents our variance calculations for absolute branch lengths and for ratios of branch lengths in this tree. Variances are slightly greater for internode branchs than for node–tip branches. Ratios of node–tip branches on opposite sides of the tree have lowest variance (their statistical correlation is less negative). Compared to a three–sequence tree (Figure 1), all variances are greater. The optimum sequence divergence also shifts to lower values (ca. 18–22%, depending upon which pair of branches are examined) and the range of the effective optimum is also smaller (ca. 12%–25%). Thus, as trees become more complex, optimal branch lengths decrease and the range of their optimum becomes narrower.

We also examined the effect of increasing one branch, an "outgroup", as was done in Figure 2. Increasing one branch causes the variances of adjacent branches to increase as in Figure 2, but the variances of branches separated by the internodal branch increase little. Thus, increased outgroup distance appears to have a largely local effect, in that only branches directly connected show increased variance.

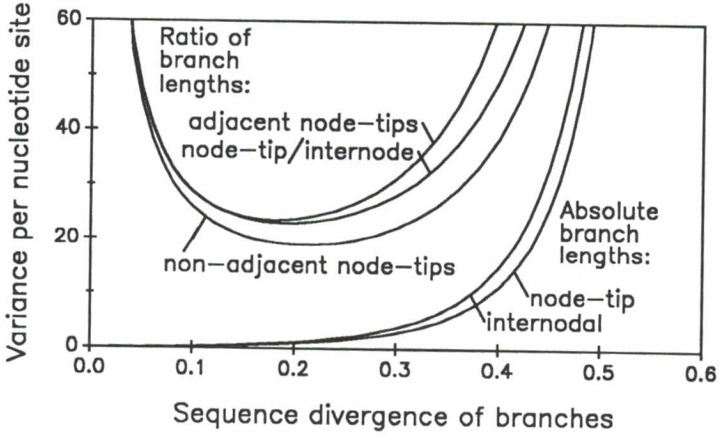


FIGURE 3. Variances for a four–sequence tree.

More complex trees involving five, six or more sequences can be investigated using the methods employed in this study, but computer time required will grow rapidly. One source of error that arises in more complex topologies, not included in the likelihood framework, is that caused by node misplacement (i.e., estimation of incorrect topology). We have assumed the correct topology is found by searching among alternative topologies and choosing the most likely (6). Although the bootstrap method can empirically determine the "variance among topologies" (5), no analytical methods exist to study this aspect of statistical error.

It is possible that some of the results reported in this investigation would be altered if this between-tree variance was included. In particular, the effect of "outgroup" taxa on the entire tree might be greater. In a previous study, the variance-effect of including distant sequences (found via the bootstrap method), suggested a "microscope analogy", in which the resolution of a phylogeny is analogous to looking at an object through a microscope of fixed angular resolution (7). Inclusion of very distant sequences forces a decrease of magnification to view the entire phylogeny, causing blurring of local small-scale topology, and subsequent misinference. The effect of this variation of sequence relatedness upon the statistical reliability of phylogeny remains an open issue.

REFERENCES

1. Crow J, Kimura M (1970)."An Introduction to Population Genetics Theory," Minneapolis: Burgess, p.509-514.
2. Nei M (1987). "Molecular Evolutionary Genetics". New York: Columbia University Press.
3. Brown WM (1985). The mitochondrial genome of animals. In RJ MacIntyre (ed): "Molecular Evolutionary Genetics", New York: Plenum, pp. 95-130.
4. Wolfe KH, Li W-H, Sharp PM (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc Nat Acad Sci USA 84: 9054-9058.
5. Felsenstein, J (1988).Phylogenies from molecular sequences: inference and reliability. Ann Rev Genet 22: 521-565.
6. Felsenstein, J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.
7. Ritland K, Clegg M (1987). Evolutionary analyses of plant DNA sequences. Amer Nat 130: S74-S100.