Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites[1]

Ziheng Yang

Department of Animal Science, Beijing Agricultural University

Felsenstein's maximum-likelihood approach for inferring phylogeny from DNA sequences assumes that the rate of nucleotide substitution is constant over different nucleotide sites. This assumption is sometimes unrealistic, as has been revealed by analysis of real sequence data. In the present paper Felsenstein's method is extended to the case where substitution rates over sites are described by the Γ distribution. A numerical example is presented to show that the method fits the data better than do previous models.

Introduction

Felsenstein (1981) presented a maximum-likelihood method for inferring phylogeny from homologous DNA sequences. This method has a firm statistical basis (Felsenstein 1981; Goldman 1990) and is powerful in recovering correct tree topologies in computer simulation studies (Fukami-Kobayashi and Tateno 1991; Hasegawa et al. 1991). Felsenstein's method, however, assumes that the rate of substitution is the same at different nucleotide sites. This assumption is unrealistic, and accumulated evidence of rate variation over sites is now overwhelming (e.g., Fitch and Margoliash 1967; Uzzell and Corbin 1971; Holmquist et al. 1983; Fitch 1986).

One attempt to take into account such spatial rate variation is to assume that some sites are invariable while all others evolve at a single rate (Hasegawa et al. 1985; Hasegawa and Kishino 1989). However, as noted by those authors, there should be a continuum of variability of sites in real sequences. Using globin-gene sequence data, Yang (1992) compared several continuous distributions for this purpose and suggested the Γ distribution as an adequate approximation. The Γ distribution has been used in constructing estimates of sequence divergence by Nei and Gojobori (1986), Jin and Nei (1990), and Li et al. (1990).

In the present paper a maximum-likelihood approach for phylogenetic inference will be presented, under the assumption that substitution rates over sites follow the Γ distribution. Other assumptions are the same as those in Felsenstein (1981), such as the independence of nucleotide substitutions at different sites and possible variation of substitution rates along different lineages.

Theory

We will derive the likelihood function for an example tree topology as shown in figure 1, by using a simple model of nucleotide substitution (Felsenstein 1981). The

[1] Key words: phylogeny inference, maximum likelihood, spatial rate variation, Γ distribution.

Address for correspondence and reprints: Dr. Ziheng Yang, Department of Zoology, Downing Street, Cambridge CB2 3EJ, United Kingdom.

Mol. Biol. Evol. 10(6):1396–1401. 1993.

0737-4038/93/1006-0019$02.00

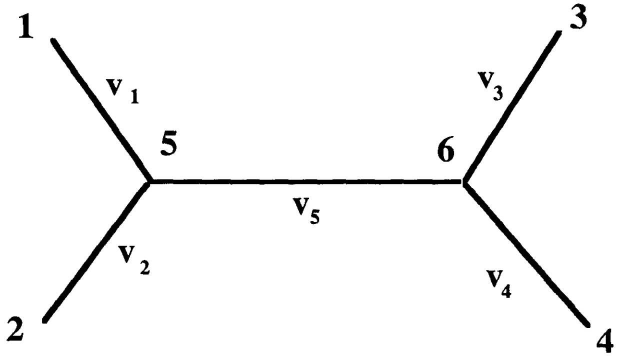

FIG. 1.—Example tree of four species used for deriving the likelihood. The tree is unrooted, and the branch lengths, v's, are measured as the average number of nucleotide substitutions per site.

extension to arbitrary tree topologies or to more complex substitution models is obvious.

According to Felsenstein's (1981) substitution model, the probability that nucleotide i changes to nucleotide j ($i \neq j$) in an infinitesimal time interval Δt is $\lambda\pi_j\Delta t$, where π_j is the equilibrium frequency of nucleotide j, and i and j take values 1, 2, 3, or 4 corresponding to T, C, A, or G, respectively. The transition probability from nucleotide i to j in relative time v ($=\lambda t$) is given by Felsenstein (1981) as

$$P_{ij}(v) = \pi_j + (\delta_{ij} - \pi_j)e^{-v} . \quad (1)$$

A rate factor, r_h, will be assigned to site h, while the overall rate is λ as before. We assume that r_h's are independently and identically distributed according to the Γ distribution

$$f(r) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta r}r^{\alpha-1}, \quad r > 0 , \quad (2)$$

with mean $E(r) = \alpha/\beta$ and variance $\mathrm{Var}(r) = \alpha/\beta^2$. β is a trivial scale factor, and to avoid the use of too many parameters we restrict the mean of the distribution to be 1 and set $\beta = \alpha$. Thus the Γ distribution is related to a single parameter α, which determines the extent of rate variation. A small α suggests that rates differ significantly over sites, while a very large α means roughly equal rates.

For the tree in figure 1 there are four sequences (species). Suppose that the length of the sequence is n. Let $\mathbf{x}_h = \{x_1, x_2, x_3, x_4\}^{\mathrm{T}}$ be the observed nucleotides at site h in the four sequences ($h = 1, 2, \ldots, n$), where the superscript "T" denotes transpose. Here we write $x_{1h}, x_{2h}, \ldots,$ as $x_1, x_2, \ldots,$ for simplicity. We also use x_5 and x_6 to designate possible nucleotides at internal nodes 5 and 6, although they are not observed. Let $\mathbf{v} = \{v_1, v_2, \ldots, v_5\}^{\mathrm{T}}$ be the five branch lengths in the tree. For data of site h, we multiply v_j ($j = 1, 2, \ldots, 5$) by a rate factor r_h.

Because of the reversibility of the substitution scheme, we can take any internal node, say node 5 in figure 1, as the ancestor (Felsenstein 1981). The conditional probability of observing $\mathbf{x}_h$, given that the rate factor of site h is r_h, will be

$$P(\mathbf{x}_h;\ \mathbf{v}\,|\,r_h) = \sum_{x_5=1}^{4}\sum_{x_6=1}^{4} \pi_{x_5}\cdot[P_{x_5x_1}(v_1r_h)\cdot P_{x_5x_2}(v_2r_h)\cdot P_{x_5x_6}(v_5r_h) \times P_{x_6x_3}(v_3r_h)\cdot P_{x_6x_4}(v_4r_h)] = \sum_{x_5=1}^{4}\sum_{x_6=1}^{4} \pi_{x_5}\cdot Y\,, \tag{3}$$

where Y is the product of the five transition probabilities in the square bracket. The unconditional probability of observing $\mathbf{x}_h$ is thus

$$P(\mathbf{x}_h;\ \mathbf{v},\,\alpha) = E\{P(\mathbf{x}_h;\ \mathbf{v}\,|\,r_h)\} = \sum_{x_5=1}^{4}\sum_{x_6=1}^{4} \pi_{x_5}\cdot E(Y)\,. \tag{4}$$

The expectation $E(\cdot)$ is taken over the random variable r_h, and an explicit form of $E(Y)$ is given in the Appendix. The log likelihood is now obtained as

$$l = \sum_{h=1}^{n} \log\{P(\mathbf{x}_h;\ \mathbf{v},\,\alpha)\} = \sum_{h=1}^{n} \log\left\{\sum_{x_5=1}^{4}\sum_{x_6=1}^{4} \pi_{x_5}\cdot E(Y)\right\}. \tag{5}$$

To maximize l over $\mathbf{v}$ and α, a nonlinear programming algorithm can be used (e.g., Gotfried and Weisman 1973, pp. 84–112). The iteration is stopped when changes in l and in $\mathbf{v}$ and α are small enough and when the gradient of the likelihood function is sufficiently close to zero. The same process is repeated for other tree topologies, and the tree with the highest likelihood is chosen as the maximum-likelihood tree (Felsenstein 1981).

A Numerical Example

As an example, we apply this approach to the α- plus β-globin gene sequence data to infer the branching order of Primates (*P*, human), Artiodactyla (*A*, goat for the α-globin gene and cow for the β-globin gene), Lagomorpha (*L*, rabbit), and Rodentia (*R*, rat). Only data from the first and second codon positions are used (570 nucleotides total), and the model of nucleotide substitution of Felsenstein (1981) is adopted.

The likelihood values and estimates of the α parameter are given in table 1 for the four possible unrooted trees. The length of the internal branch in tree 3 approaches zero, and thus tree 3 approaches tree 1. The best tree is tree 2, with the branching order and branch lengths to be (*P*, 0.050; *A*, 0.106; central branch, 0.037; *L*, 0.063; and *R*, 0.185) with $\hat{\alpha}$ = 0.286. To test the significance of tree 2, we calculate the bootstrap probabilities by using the approximate method of Kishino et al. (1990), which resamples the estimated log likelihoods (the RELL method) at each site. Table 1 shows that the maximum-likelihood tree is not significantly supported by this test.

The models' adequacy can be compared by using their likelihood values calculated from the same data. Felsenstein's (1981) method, which assumes constant rates over sites, gives the same best tree, with log likelihood of −1453.18. The improvement in log likelihood gained by assuming a Γ distribution of rates over sites is (−1436.65) − (−1453.18) = 16.52. The constant rate over sites is the limiting case of the Γ distribution when α approaches infinity. So we may use the likelihood-ratio test to compare the two models; we compare $2\Delta l$ = 33.04 with $\chi^2_{0.01}$ = 6.63 (df = 1), and the difference is seen to be significant.

Table 1
Comparison of Different Tree Topologies among Primates (P), Artiodactyla (A), Lagomorpha (L), and Rodentia (R)[a]

Tree	l_i	$l_i - l_2$	$\hat{\alpha}$[b]	P[c] (%)
1. (P, A, L, R)	−1,440.69	−4.04	0.223 ± 0.061	0.2
2. ((P, A), L, R) ...	−1,436.65	0.00	0.286 ± 0.087	77.6
3. ((P, L), A, R) ...		same as tree 1		
4. ((P, R), L, A) ...	−1,439.43	−2.78	0.246 ± 0.071	22.2

[a] The first- and second-codon-position data of the α- and β-globin genes are used.
[b] Standard errors of $\hat{\alpha}$ are calculated by the curvature method.
[c] P is the bootstrap probability, calculated by the RELL method (Kishino et al., 1990). See text for more comments.

Assuming the constant rate over sites but adopting Hasegawa et al.'s (1985) substitution scheme also produces the same best tree, with log likelihood of −1451.01. The estimated transition/transversion ratio is 1.48, and Felsenstein's (1981) substitution scheme is not seriously in error.

The Computer Program

A C program that implements the method described in this paper is available from the author. It is very slow, and computation time increases explosively with the number of species. With a microcomputer, it seems impractical to compare tree topologies with more than four species.

Discussion

The α parameter need not be estimated by iteration for each of the tree topologies. The estimate is rather insensitive to topologies, and therefore the estimate from the star tree, which involves much less computation, can be used in later calculations. Estimates from the parsimony method are also usable (Holmquist et al. 1983; Kocher and Wilson 1991). When α is found this way, the speed of the algorithm can be improved considerably.

As the present approach involves much more computation than that of Felsenstein (1981), it is worthwhile to examine how Felsenstein's method behaves in the presence of spatial rate variation. According to Jin and Nei's (1990) simulations, the neighbor-joining method might perform poorly if spatial rate variation is not taken into account in the estimation of sequence divergence. It is not clear whether the same conclusion can apply for the maximum-likelihood method.

Acknowledgments

Professors Changhsin Wu and Yuan Chang gave much encouragement while I performed this study. Dr. Adrian Friday provided facilities for the revision of the manuscript, and Dr. Nick Goldman gave many useful suggestions for improvement. I also wish to thank Professor M. Nei and two anonymous reviewers for many critical comments on the manuscript.

APPENDIX

Derivation of $E(Y)$

Equation (1) can be written as

$$P_{ij}(v) = C_{ij0} + C_{ij1}e^{-v}, \tag{A1}$$

where $C_{ij0} = \pi_j$ and $C_{ij1} = (\delta_{ij} - \pi_j)$. Y in eq. (3) is a product of five terms, each of the form of eq. (A1). To obtain $E(Y)$, we expand Y into the sum of 2^5 terms: $E(Y)$ will then be the sum of the expectation of each of these terms. If we let M_{mj} be the jth bit (0 or 1) when $(m - 1)$ is expressed as a binary number, we obtain $E(Y)$ as

$$E(Y) = \sum_{m=1}^{2^5} [C_{x_5x_1M_{m1}} \cdot C_{x_5x_2M_{m2}} \cdot C_{x_5x_6M_{m5}} \cdot C_{x_6x_3M_{m3}} \cdot C_{x_6x_4M_{m4}}] \cdot \left[\frac{\alpha}{\alpha + S}\right]^{\alpha}, \tag{A2}$$

where $S = \sum_{j=1}^{5} M_{mj}v_j$.

LITERATURE CITED

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

FITCH, W. M. 1986. The estimate of total nucleotide substitutions from pairwise differences is biased. Philos. Trans. R. Soc. Lond. [Biol.] **312**:317–324.

FITCH, W. M., and E. MARGOLISH. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem. Genet. **1**:65–71.

FUKAMI-KOBAYASHI, K., and Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitution. J. Mol. Evol. **32**:79–91.

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.

GOTFRIED, B. S., and J. WEISMAN. 1973. Introduction to optimization theory. Prentice-Hall, Englewood Cliffs, N.J.

HASEGAWA, M., and H. KISHINO. 1989. Confidence limits on the maximum likelihood estimation of the hominoid tree from mitochondrial DNA sequences. Evolution **43**:672–677.

HASEGAWA, M., H. KISHINO, and N. SAITOU. 1991. On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. **32**:443–445.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HOLMQUIST, R., M. GOODMAN, T. CONRY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. **19**:137–448.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogeny analysis. Mol. Biol. Evol. **7**:82–102.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. Pp. 391–413 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer, Tokyo.

LI, W.-H., M. GOUY, P. M. SHARP, C. O'HUIGIN, and Y.-W. YANG. 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla and Carnivora and molecular clocks. Proc. Natl. Acad. Sci. USA **87**:6703–6707.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172:**1089–1096.

YANG, Z. 1992. Variations of substitution rates and estimation of evolutionary distances of DNA sequences. Ph.D. thesis, Beijing Agricultural University, Beijing.

MASATOSHI NEI, reviewing editor

Received October 1, 1992; revision received April 29, 1993

Accepted May 17, 1993