J Mol Evol (1994) 38:305–309

Journal of Molecular Evolution

Monte Carlo Simulation in Phylogenies: An Application to Test the Constancy of Evolutionary Rates

José Carlos Adell,[1] Joaquín Dopazo[2]*

[1] Departament de Genètica, Universitat de València, C/Dr. Moliner 50, E-46100 Burjassot, València, Spain
[2] Centro de Investigación en Sanidad Animal, INIA, E-28130, Valdeolmos, Madrid, Spain

Received: 26 March 1993 / Revised: 27 July 1993

Abstract. Monte Carlo simulation has commonly been used in phylogenetic studies to test different tree-reconstruction methods, and consequently, its application for testing evolutionary models can be considered as a natural extension of this usage. Repetitive simulation of a given evolutionary process, under the restrictions imposed by the model to be tested, along a determinate tree topology allow the estimate of probability distributions for the desired parameters. Next, the phylogenetic tree can be reconstructed again without the constraints of the model, and the parameter of interest, derived from this tree, can be compared to the corresponding probability distribution derived from the restricted, simulated trees. As an example we have used Monte Carlo simulation to test the constancy of evolutionary rates in a set of cytochrome-c protein sequences.

Key words: Monte Carlo simulation — Parametric bootstrap — Molecular clock — Evolutionary rates — Phylogeny — Least-squares method — Cytochrome-c

Introduction

Computer-intensive methods based on resampling (Efron and Tibshirani 1991), have recently gained importance in the field of phylogenetic analysis (Felsenstein 1988) because they require very little in the way of modeling, assumptions, or analysis, and can be automatically applied to the estimation of any parameter, no matter on how complicated its actual distribution is (Efron and Gong 1983). One class of computer-intensive method is the Monte Carlo simulation, in which specific assumptions concerning the model to be simulated are required. Felsenstein (1988) proposed the use of the "parametric bootstrap" (Efron 1985) for phylogenetic analysis. This approach consists simply in taking the best tree and simulating new data sets of the same size by evolution occurring along that tree under the postulated model (Felsenstein 1988). The approximation of the "parametric bootstrap" can be used not only to evaluate the variability held by the original estimate of the tree, as proposed by Felsenstein (1988); it also can be extended to test whether the evolution of a given set of data can be described by a given model. In this approach, the tree is reconstructed under the restrictions imposed by the model and the evolution along it is simulated following the model. Next, the tree is reconstructed free of the constraints of the model and tested against the constrained tree.

To illustrate this approximation we have chosen as evolutionary model the molecular clock. Here, we show how to perform a Monte Carlo simulation of evolution under a molecular clock model and how to test whether a set of data actually fits to the expectations of the model.

Materials and Methods

Sequences of Cytochrome-c and Evolutionary Pattern. We have used a set of cytochrome-c protein sequences obtained from the NBPF-PIR database (release 23). The corresponding species were *Saccharomyces*

Correspondence to: J. Dopazo
* *Present address:* Centro Nacional de Biotecnología, CSIC, Universidad Autonoma, E-28049, Madrid, Spain

cerevisiae (accession code (AC) = A00037) and *Triticum aestivum* (AC = A00060) (used as outgroups); *Helix aspersa* (AC = A00029); *Eisenia foetida* (AC = A00027); *Macrobrachium malcolmsonii* (AC = A00028); *Manduca sexta* (AC = A00032); *Drosophila melanogaster* (AC = A00030); *Asterias rubens* (AC = A00026); *Euthynnus pelamis* (AC = A00022); *Lampetra tridentata* (AC = A00025); *Rana catesbeiana* (AC = A00021); *Columba livia* (AC = A00013); *Anas platyrhynchos* (AC = A00015); *Oryctolagus cuniculus* (AC = A0009); and *Homo sapiens* (AC = A00001). These species constitute a reasonable representation of the animal kingdom. Also *Saccharomyces cerevisiae* and *Triticum aestivum* have been included, and used as outgroups.

The branching pattern (topology of the evolutionary tree) used to relate these sequences to each other was that proposed by Margulis and Schwartz (1982).

The sequences were aligned by means of the multiple-sequence alignment algorithm developed by Higgins and Sharp (1989), using the program CLUSTAL.

Method of Phylogenetic Reconstruction for Both Expected and Observed Trees. The least-squares method (Fitch and Margoliash 1967) was used because it is available in the PHYLIP package (Felsenstein 1990) in two versions: the KITSCH program, which implements the method under the molecular clock assumption; and the FITCH program, which implements the same method with no constraints imposed for branch lengths. KITSCH was used to obtain a least-squares estimation of the "expected," clocklike, phylogenetic tree, whereas FITCH was used to obtain the "observed" one. Thus, differences between the "expected" and the "observed" trees due to the use of different tree-reconstruction methods are avoided.

Pairwise distance matrices were obtained using Kimura's modified estimator (Kimura 1983, p. 75).

Simulation of the Evolutionary Pattern. The evolutionary branching pattern simulated was the one proposed by Margulis and Schwartz (1982). The expectations for branch lengths under a molecular clock were obtained by the application of the KITSCH program (least-squares assuming a molecular clock) for the previous branching pattern.

For each simulation, an ancestral sequence was randomly generated with a length (**L**) equal to that of the aligned set of sequences and with an amino acid composition identical to the shown by the consensus sequence. The topology of the tree is traversed in preorder (Knuth 1973) from the ancestral sequence to all the tips. Each descendent sequence, **d,** was obtained from its corresponding ancestor, **a,** by changing each individual position with probability $\mathbf{P_{a,d}/L}$, where $\mathbf{P_{a,d}}$ was the expected number of mutations, under the molecular clock model, along the branch leading from sequence **a** to **d.** All possible changes between the 20 amino acids were assumed to be equiprobable. This probability of change corresponds to the "Poisson model" (see Adachi and Hasegawa 1992; Adachi et al. 1993)—the simplest and most widely used model in phylogenetic studies, even if not stated explicitly.

Test Based on the Empirical Distribution of Expected Values. The aim of the test based on the simulation was to obtain the probability distribution for all the branch lengths under the molecular clock assumption. These distributions will allow one to check whether the observed branch-length values belong to their corresponding distribution or, on the contrary, whether they can be considered significantly different from their clocklike expectations. To carry out the test, 1,000 clocklike data sets were simulated as described above and, based on them, 1,000 phylogenetic trees were reconstructed by the KITSCH program. (The tree topology was forced to be the same with the "user tree" option; see Felsenstein 1990.) For each simulated tree, all the branch-length values were recorded. From them, the distribution of values for each clocklike branch was constructed. The distributions were then used to test the observed branch-length values by means of two-tailed percentile tests.

Results

Figure 1A shows the tree obtained with the FITCH program, and Fig. 1B shows the tree obtained with the KITSCH program. Branches represented with a dotted line were zero in the case of the FITCH tree, or nonsignificantly positive (see Dopazo 1993) in the case of the KITSCH tree. Table 1 shows the "observed" unconstrained branch lengths, as well as the averages corresponding to the constant-rate tree obtained by simulation. When the "observed" values were compared to the "expected" constant-rate distributions obtained by simulation, some stretches of the tree actually revealed significant departures from their expectations under the assumption of constant-rate evolution (Table 1). Internal edges showing zero values in KITSCH that were positive in FITCH were detected as nonclocklike lineages. *E. foetida* and *H. sapiens* lineages showed a higher evolutionary rate than expected under a molecular clock.

The application of the inequality rate test (Felsenstein 1984) rendered a $F(13,78) = 3.21$, which indicates that the observed tree is significantly different from a constant-rate tree at 99.9% level.

Figure 2 shows the probability distribution obtained by simulation for four branch lengths. These include the terminal branches leading to *H. sapiens* and *E. foetida,* which showed a faster evolutionary rate than expected under the assumption of a molecular clock, the terminal branch leading to *M. malcolmsonii,* which showed a clocklike behavior, and an internal branch showing an evolutionary rate lower than expected under a molecular clock assumption (although this rate was not low enough to be considered significantly different from a constant rate).

Discussion

Although Dickerson (1971) claimed that cytochrome-c behaves as a molecular clock, our analysis, which includes a slightly different set of sequences, has shown that a molecular clock cannot be assumed for all the species studied. In fact, local deviations from a constant rate of evolution have been detected. A common procedure used to test the molecular clock hypothesis has been the dispersion index (Kimura 1968, 1983, 1987; Langley and Fitch 1974; Gillespie 1986; Gillespie and Langley 1979), although recently, Gillespie (1989) has pointed out that the clock could be strongly affected by the topology of the phylogeny or by lineage effects due to differences in generation times. A similar problem was noticed by Felsenstein (1988) for the "relative rate

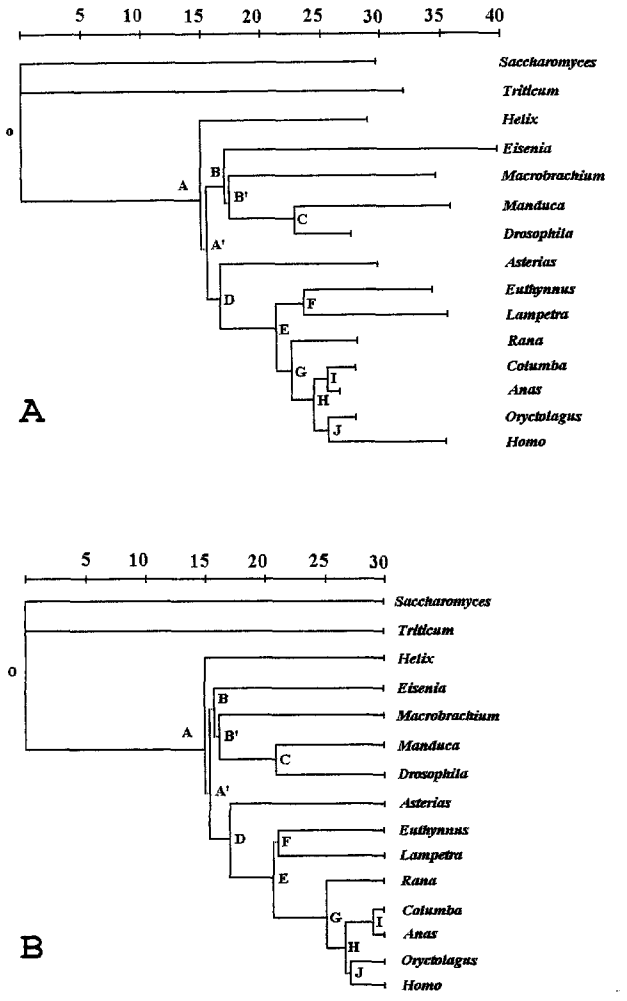

Fig. 1. Phylogenetic trees for the evolution of cytochrome-c. **A** "Observed tree" obtained using the FITCH program. **B** "Expected tree" under constant-rate assumption obtained using the KITSCH program. *Dotted lines* represent branches not significantly positive.

test" (Sarich and Wilson 1967). To avoid this problem, different tests, which compare two phylogenetic trees inferred for the same set of data, have been proposed. Usually, two trees, one of them inferred under the assumption that its evolutionary process has been directed by a molecular clock ("clocklike tree"), and the other tree inferred without such an assumption ("observed tree"), are contrasted. This is the case of the test for inequality of evolutionary rates (Felsenstein 1984), and the more sophisticated likelihood ratio test (Langley and Fitch 1974; Felsenstein 1990).

Analogous to Monte Carlo simulation, maximum-likelihood (ML) methods involve the adoption of a model for the evolution of the data. Indeed, in this paper, a model currently used in ML phylogenetic studies has been employed (Adachi and Hasegawa 1992; Adachi et al. 1993). This model provides a very simple description of the process which describes the evolution of the protein sequences because it does not take into account the different probabilities of change shown by different amino acids at different sites of the molecule (see Pesole et al. 1992). Nevertheless, in spite of its simplicity, the Poisson model has been demonstrated to be efficient in ML tree reconstruction (Adachi et al. 1993) and, consequently, it can be considered a good approach to the study of the evolutionary processes of the sequences. In ML methods, the model must commonly be extremely simple because of the complexity of the calculations; however, from the point of view of simulation, the use of more complicated models does not increase the complexity in the calculations to the same extent that a ML approach does. Moreover, evolution of DNA or protein sequences and even DNA restriction sites and other characters can be simulated, including features of difficult modeling in algebraic models (such as the correlation between positions due to the necessity of maintaining protein structure or RNA foldings, etc.), without a significant increase in the complexity of the calculations.

The modification of the "parametrical bootstrap" (Efron 1985; Felsenstein 1988) here proposed allows the expected tree (a tree inferred under the assumption of constancy in the evolutionary rates) to be used to simulate new data sets with evolution occurring along the tree under the postulated model. Thus, the variability among the trees estimated from those simulated data sets can be used to construct probability distributions for the parameter (or parameters) of interest—in this case the branch-length values. At this point, a simple two-tailed test can be used to assess whether the observed parameter falls into the expectations under the model.

Although the classical inequality rate test (Felsenstein 1984) was able to detect a significant nonconstancy in the evolutionary rates of the studied set of proteins, the simulation approach provides more information, indicating which branches significantly deviate from the assumptions of the model, as well as the trend of this deviation.

Simulation can be considered as a well-established method in the field of phylogeny although its most common use has been for testing tree-reconstruction methods. Thus, simulation for gene frequency (Astolfi et al. 1981; Nei et al. 1983; Kim and Burgman 1988), discrete morphological characters (Fiala and Sokal 1985; Sokal 1983), and molecular data (Peacock and Boulter 1975; Tateno et al. 1982; Li et al. 1987a,b; Tateno and Tajima 1986; Saitou 1988; Rzhetsky and Nei 1992), to cite just a few cases, has been extensively used. Hence, the method here proposed constitutes a natural extension of the application of the simulation to phylogenetic studies.

Since the simulation allows the comparison of observed values to their expectations under some assumption of interest, other applications besides testing evolutionary rates can be performed within this framework. So, comparison between different types of nucleotide or amino acid substitutions can be performed, providing that the expected values can be modeled by

Table 1. Observed number of mutations (obtained by FITCH program), expected number of mutations under constant rate assumption (KITSCH program), average branch lengths and standard deviations obtained from 1,000 simulated replicates of the constant rate tree

Branch	FITCH	KITSCH	Branch length	SD
Origin—A	15.43	15.87	16.01	4.12
Origin—*Triticum aestivum*	31.20	30.12	30.14	4.72
Origin—*Saccharomyces cerevisiae*	29.05	30.12	30.15	4.63
A—*Helix aspersa*	13.32	14.54	14.57	3.84
A—B	1.68[a]	0.00	0.00	0.00
B—*Eisenia foetida*	21.75[a]	14.54	14.58	3.62
B—*Macrobrachium malcolmsonii*	16.76	14.54	14.51	3.51
B—C	4.96	5.24	5.16	2.19
C—*Manduca sexta*	12.59	9.30	9.32	3.07
C—*Drosophila melanogaster*	6.02	9.30	9.32	2.90
A—D	0.12	1.53	1.54	1.25
D—*Asterias rubens*	13.01	13.00	13.01	3.48
D—E	4.68	3.12	3.12	1.77
E—F	1.94[a]	0.00	0.00	0.00
F—*Euthynnus pelamis*	10.23	9.90	9.87	2.96
F—*Lampetra tridentata*	10.87	9.90	10.11	2.99
E—G	0.56	3.46	3.47	1.84
G—*Rana catesbeiana*	6.52	6.40	6.46	2.38
G—H	2.36	2.42	2.35	1.53
H—I	2.05	2.48	2.50	1.55
I—*Columba livia*	1.99	1.52	1.51	1.25
I—*Anas platyrhinchos*	1.06	1.52	1.48	1.23
H—J	1.26[a]	0.00	0.00	0.00
J—*Oryctolagus cuniculus*	1.81	4.00	3.94	1.88
J—*Homo sapiens*	7.63[a]	4.00	3.96	1.93

[a] Indicates values of the branches significantly different ($\alpha = 0.05$) from their corresponding constant-rate expectations obtained by simulation

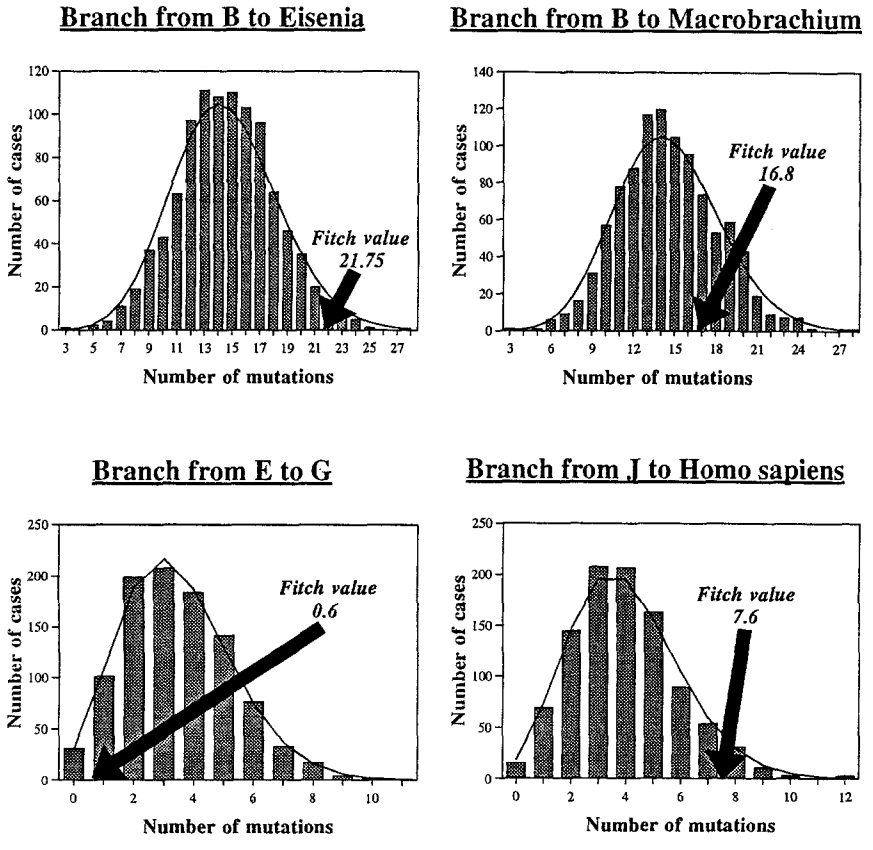

Fig. 2. Empirical distributions for clocklike branch-length values obtained by simulation. The *continuous line* represents the theoretical Poisson distribution and the *arrow* shows the observed branch-length value obtained upon the application of the FITCH program. See text for discussion.

simulation. Also, under some circumstances and for some cases, a test similar to this one can be used in a "comparative method"—i.e., to compare characters across species so as to discover patterns in character evolution (Ridley 1983; Felsenstein 1985; Donoghue 1989; Maddison 1990).

Computer-intensive methods such as the approach proposed here, based on Monte Carlo simulation, provide a very attractive alternative to ML methods (see Efron 1985) in statistical studies of evolutionary processes in sequences.

Acknowledgments. The authors thank F. Sobrino, J.L. Ménsua, F. Gonzalef-Candelas, and A. von Haeseler for their valuable comments, and J. Felsenstein, who kindly provided his PHYLIP package of programs. Work at INIA was supported by grant 9022 from INIA and by Tecnologia para Diagnostico e Investigación S.A.

References

Adachi J, Cao Y, Hasegawa M (1993) Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J Mol Evol 36:270–281

Adachi J, Hasegawa M (1992) Amino acid substitution of proteins coded for in mitochondrial DNA during mammalian evolution. Jpn J Genet 67:187–197

Astolfi P, Kidd KK, Cavalli-Sforza LL (1981) A comparison of methods for reconstructing evolutionary trees. Syst Zool 30:156–169

Dickerson RE (1971) The structure of cytochrome-c and the rates of molecular evolution. J Mol Evol 1:26–45

Donoghue MJ (1989) Phylogenies and the analysis of evolutionary sequences, with examples from seed plants. Evolution 43:1137–1156

Dopazo J (1993) Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. J Mol Evol In press

Efron B (1985) Bootstrap confidence intervals for a class of parametric problems. Biometrika 72:45–58

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Am Stat 37:36–48

Efron B, Tibshirani R (1991) Statistical data analysis in the computer age. Science 253:390–395

Felsenstein J (1984) Distance methods for inferring phylogenies: a justification. Evolution 38:16–24

Felsenstein J (1985) Phylogenies and the comparative method. Am Natur 125:1–15

Felsenstein J (1988) Phylogenies from molecular sequences: inferences and reliability. Annu Rev Genet 22:521–565

Felsenstein J (1990) PHYLIP Manual version 3.3. University Herbarium, University of California, Berkeley, CA

Fiala KL, Sokal RR (1985) Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution 39:609–622

Fitch WM, Margoliash E (1967) The construction of phylogenetic trees. Science 155:279–284

Gillespie JH (1986) Variability of evolutionary rates of DNA. Genetics 113:1077–1091

Gillespie JH (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol 6:636–647

Gillespie JH, Langley CH (1979) Are evolutionary rates really variable? J Mol Evol 13:27–34

Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. Comput Appl Biosci 5:151–153

Kim J, Burgman MA (1988) Accuracy of phylogenetic estimation methods using simulated allele-frequency data. Evolution 42: 596–602

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kimura M (1987) Molecular evolutionary tree and the neutral theory. J Mol Evol 26:24–33

Knuth DE (1973) Fundamental algorithms. Addison-Wesley, Reading, MA

Langley CH, Fitch WM (1974) An estimation of the constancy of the rate of molecular evolution. J Mol Evol 3:161–177

Li W-H, Tanimura M, Sharp PM (1987a) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Li W-H, Wolfe KH, Sourdis J, Sharp PM (1987b) Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harbor Symp Quant Biol 52:847–856

Madison WP (1990) A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44:539–557

Margulis L, Schwartz KV (1982) Five kingdoms. An illustrated guide to the phyla of life on hearth. Freeman Co, New York

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees for molecular data. II. Gene frequency data. J Mol Evol 19:153–170

Peacock D, Boulter D (1975) Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J Mol Biol 95:513–527

Pesole G, Attimonelli M, Preparata G, Saccone C (1992) A statistical method for detecting regions with different evolutionary dynamics in multialigned sequences. Mol Phylogen Evol 1:91–96

Ridley M (1983) The explanation of organic diversity. The comparative method and adaptations of mating. Clarendon, Oxford

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9:945–967

Saitou N (1988) Property and efficiency of the maximum-likelihood method for molecular phylogeny. J Mol Evol 27:261–273

Sarich VM, Wilson AC (1967) Immunological time scale for hominoid evolution. Science 158:1200–1203

Sokal RR (1983) A phylogenetic analysis of the Caminalcules. II. Estimation of the true cladogram. Syst Zool 32:185–201

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related trees. J Mol Evol 18:387–404

Tateno Y, Tajima F (1986) Statistical properties of molecular tree construction methods under the neutral mutation model. J Mol Evol 23:353–361