Phylogenetics and Molecular Evolution/Filogenética e Evolução Molecular

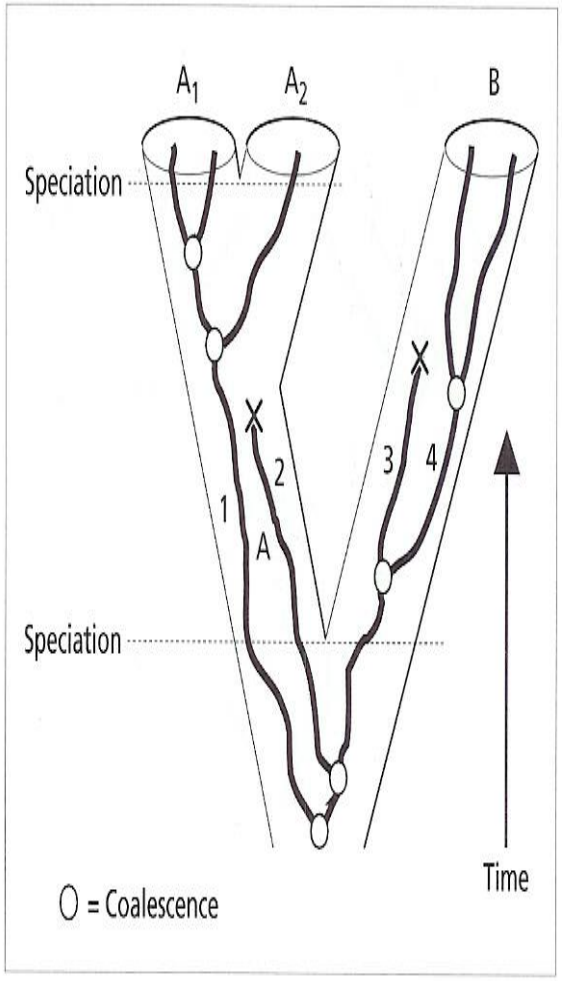
Octávio S. Paulo
Computational Biology and Population Genomics Group (CoBiG2)

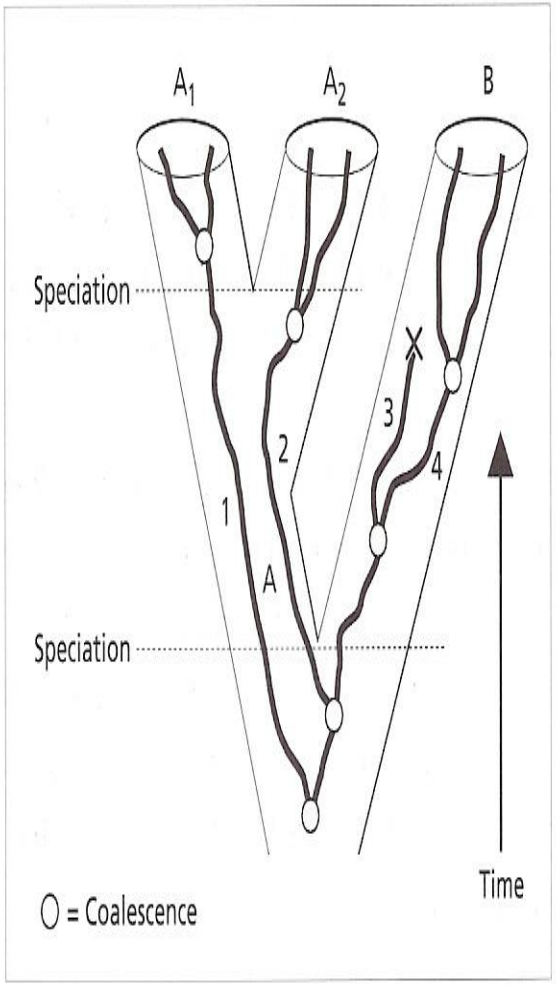
Análise de matrizes de dados com múltiplas partições

Sumário

A análise de matrizes de dados com múltiplas partições. O uso simultâneo de diferentes tipos de dados. “Gene trees” e “Species Trees”. Teste de hipótese e robustez.

Árvores de genes e de espécies podem contar histórias diferentes





Incongruence

Two phylogenies inferred from different genes can be incongruent for three reasons:

(i) stochastic error, which results from the fact that, when a limited number of characters is available, a few positions biased by multiple substitutions (i.e. convergence or reversion) can, by chance, dominate and lead to an erroneous tree;

(ii) the departure of the gene phylogeny from the species phylogeny, which can be due to undetected gene duplication (i.e. hidden paralogy), lineage sorting of multiple alleles, horizontal gene transfer or gene conversion;

(iii) systematic error, which is due to the inaccuracy of the methods of tree reconstruction used.

Philipe and Telford 2006 TREE

Incongruence - systematic errors

The known phenomena generating systematic errors can be classified into four main categories:

(i) variable nucleotide and/or amino-acid composition across taxa (i.e. the same nucleotide is independently acquired by distantly related species because the G+C content of their genomes is similar);
(ii) reduced number of possible amino acids at a given position (thereby increasing the probability of independent acquisition of the same nucleotide);
(iii) variable evolutionary rate inside sites (i.e. heterotachy);
(iv) non-independence of positions owing to structural constraints.

Philipe and Telford 2006 TREE

Glossary

Heterotachy: refers to the fact that the evolutionary rate of a given position varies throughout time. Fitch proposed the covarion model to explain this property.

Orthologous: homologous genes in two of more organisms that are related only by lineage splitting and not by gene duplication

Tests

Homogeneity of base composition

Bootstrap
Parametric Bootstrap
Ultrafast Bootstrap Approximation
SH-aLRT

LRT – Likelihood Ratio Test

Kishino-Hasegawa test (Kishino and Hasegawa 1989)
Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999)
AU test (Shimodaira, 2002)

ILD test
Likelihood Heterogeneity Test

Homogeneity of base composition

Output Display

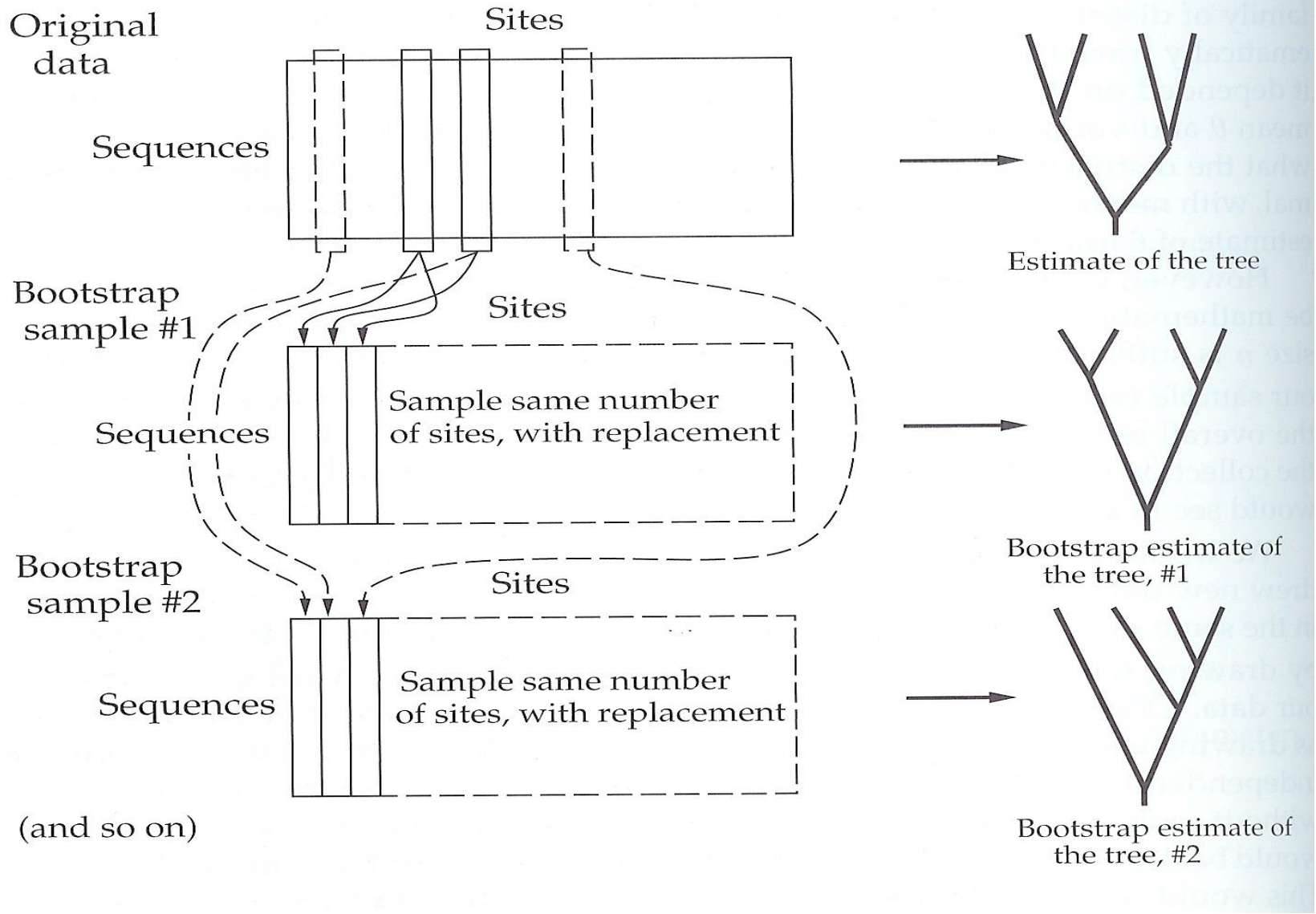
```
-----------------------------------------------------------------------
Mean          0.29118     0.29871     0.13407     0.27604  1140.00

Chi-squared test of homogeneity of state frequencies across taxa:

Taxon               A       C       G       T
-----------------------------------------------
Plotor      O   348.00  309.00  147.00  336.00
            E   331.95  340.53  152.84  314.68
Elutris     O   334.00  347.00  156.00  303.00
            E   331.95  340.53  152.84  314.68
Ggulo       O   322.00  344.00  166.00  308.00
            E   331.95  340.53  152.84  314.68
Maltaica    O   328.00  341.00  150.00  321.00
            E   331.95  340.53  152.84  314.68
Mamerican   O   319.00  357.00  165.00  299.00
            E   331.95  340.53  152.84  314.68
Merminea    O   342.00  348.00  141.00  309.00
            E   331.95  340.53  152.84  314.68
Meversman   O   333.00  342.00  148.00  317.00
            E   331.95  340.53  152.84  314.68
Mflavigul   O   329.00  340.00  160.00  311.00
            E   331.95  340.53  152.84  314.68
Mfoina      O   315.00  360.00  166.00  299.00
            E   331.95  340.53  152.84  314.68
Mfuro       O   334.00  342.00  148.00  316.00
            E   331.95  340.53  152.84  314.68
Mitatsi     O   341.00  334.00  142.00  323.00
            E   331.95  340.53  152.84  314.68
Mlutreola   O   333.00  343.00  148.00  316.00
            E   331.95  340.53  152.84  314.68
Mmartes     O   320.00  353.00  168.00  299.00
            E   331.95  340.53  152.84  314.68
Mmelampus   O   325.00  357.00  162.00  296.00
            E   331.95  340.53  152.84  314.68
Mmelesmel   O   346.00  323.00  145.00  326.00
            E   331.95  340.53  152.84  314.68
Mnivalis    O   327.00  339.00  154.00  320.00
            E   331.95  340.53  152.84  314.68
Mputorius   O   336.00  341.00  146.00  317.00
            E   331.95  340.53  152.84  314.68
Msibirica   O   335.00  342.00  144.00  319.00
            E   331.95  340.53  152.84  314.68
Mvison      O   340.00  308.00  148.00  344.00
            E   331.95  340.53  152.84  314.68

Chi-square = 33.992727 (df=54), P = 0.98486245
Warning: This test ignores correlation due to phylogenetic structure.
```

bootstrap

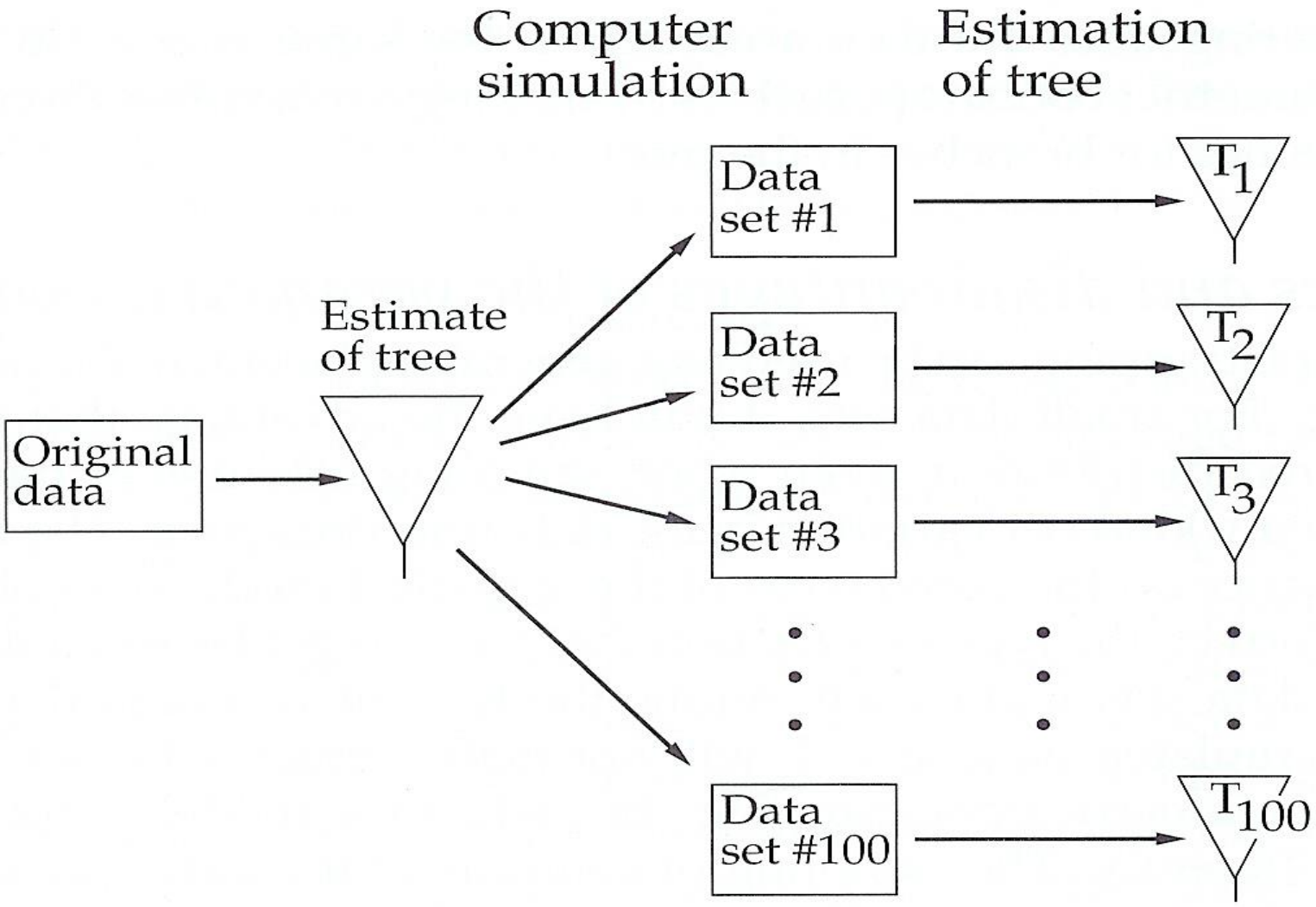

Original data
Sites
Sequences
Estimate of the tree
Bootstrap sample #1
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #1
Bootstrap sample #2
Sites
Sequences
Sample same number of sites, with replacement
Bootstrap estimate of the tree, #2
(and so on)

bootstrap

Three resample techniques are sometimes used to assess the robustness of branches within trees, nonparametric bootstrapping, jacknife and parametric bootstrapping. Only nonparametric bootstrapping, simply called bootstrapping, because it is widely used, is introduced here. The pseudoreplicate data sets are generated by randomly sampling with reposition the original character matrices of the same size as the original (Felsenstein 1985). The frequency with which a given branch is found upon analysis of these pseudoreplicate data sets is recorded as the bootstrapping proportion. These proportions are used to assess the reliability of individual branches in the optimal tree (Hillis *et al.* 1996b).

There are two important caveats related with the bootstrap technique. The first one is that it assumes that each site is independent and that there is a single distribution of rate of evolutionary change across all sites, which, at least for mtDNA is not usually the case. The second caveat is that bootstrap results are usually summarised by a majority-rule consensus tree, and if there are sequences that "float" over the trees, i.e., sequences that appear in several positions of the bootstrap trees, they lower the bootstrap value of those parts where they appear, and consequently otherwise robust parts become weakly supported. As a result of this, but not only this, it can underestimate branches with high support and overestimate the confidence of the ones with low support (Felsenstein 1985; Li & Zharkikh 1994; 1995).
Finally it is important to realise that bootstrap values give an indication of the precision of the results, not of their accuracy. Wrong models can generate wrong trees but with robust bootstrap support.

Parametric bootstrap



UFBoot2: Improving the Ultrafast Bootstrap Approximation

Diep Thi Hoang,[†,1] Olga Chernomor,[†,2] Arndt von Haeseler,[2,3] Bui Quang Minh,[*,2] and Le Sy Vinh[*,1]

[1]Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

[2]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Vienna, Austria

[3]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

[†]These authors contributed equally to this work.

[*]**Corresponding authors:** E-mails: minh.bui@univie.ac.at; vinhls@vnu.edu.vn.

Associate editor: Michael S. Rosenberg

Abstract

The standard bootstrap (SBS), despite being computationally intensive, is widely used in maximum likelihood phylogenetic analyses. We recently proposed the ultrafast bootstrap approximation (UFBoot) to reduce computing time while achieving more unbiased branch supports than SBS under mild model violations. UFBoot has been steadily adopted as an efficient alternative to SBS and other bootstrap approaches. Here, we present UFBoot2, which substantially accelerates UFBoot and reduces the risk of overestimating branch supports due to polytomies or severe model violations. Additionally, UFBoot2 provides suitable bootstrap resampling strategies for phylogenomic data. UFBoot2 is 778 times (median) faster than SBS and 8.4 times (median) faster than RAxML rapid bootstrap on tested data sets. UFBoot2 is implemented in the IQ-TREE software package version 1.6 and freely available at http://www.iqtree.org.

Key words: **phylogenetic inference, ultrafast bootstrap, maximum likelihood, model violation, polytomies.**

SH-aLRT

Approximate likelihood-ratio test relies on the Nonparametric Shimodaira-Hasegawa-like like procedure

Syst. Biol. 59(3):307–321, 2010

DOI:10.1093/sysbio/syq010
Advance Access publication on March 29, 2010

New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0

STÉPHANE GUINDON[1,2], JEAN-FRANÇOIS DUFAYARD[1], VINCENT LEFORT[1], MARIA ANISIMOVA[1,3,4], WIM HORDIJK[1,5], AND OLIVIER GASCUEL[1,*]

[1]*Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Université de Montpellier, 34392 Montpellier Cedex 5, France;* [2]*Department of Statistics, University of Auckland, Auckland 1142, New Zealand;* [3]*Institute of Computational Science, ETH, CH-8092 Zurich, Switzerland;* [4]*Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; and* [5]*Department of Statistics, University of Oxford, OX1 3TG Oxford, UK;* **Correspondence to be sent to: Olivier Gascuel, Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Université de Montpellier, 161 rue Ada, 34392 Montpellier Cedex 5, France; E-mail: gascuel@lirmm.fr.*

Received 26 June 2008; reviews returned 18 August 2008; accepted 20 December 2009
Associate Editor: Susanne S. Renner

Abstract.—PhyML is a phylogeny software based on the maximum-likelihood principle. Early PhyML versions used a fast algorithm performing nearest neighbor interchanges to improve a reasonable starting tree topology. Since the original publication (Guindon S., Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704), PhyML has been widely used (>2500 citations in ISI Web of Science) because of its simplicity and a fair compromise between accuracy and speed. In the meantime, research around PhyML has continued, and this article describes the new algorithms and methods implemented in the program. First, we introduce a new algorithm to search the tree space with user-defined intensity using subtree pruning and regrafting topological moves. The parsimony criterion is used here to filter out the least promising topology modifications with respect to the likelihood function. The analysis of a large collection of real nucleotide and amino acid data sets of various sizes demonstrates the good performance of this method. Second, we describe a new test to assess the support of the data for internal branches of a phylogeny. This approach extends the recently proposed approximate likelihood-ratio test and relies on a nonparametric, Shimodaira–Hasegawa–like procedure. A detailed analysis of real alignments sheds light on the links between this new approach and the more classical nonparametric bootstrap method. Overall, our tests show that the last version (3.0) of PhyML is fast, accurate, stable, and ready to use. A Web server and binary files are available from http://www.atgc-montpellier.fr/phyml/. [Bootstrap analysis; branch testing; LRT and aLRT; maximum likelihood; NNI; phylogenetic software; SPR; tree search algorithms.]

The likelihood ratio test statistic (Huelsenbeck & Crandall 1997)

Models are of critical importance to estimate the rate of evolution, divergence time and to reconstruct phylogenetic trees, so several tests have been developed to assess the best hypotheses. The Likelihood ratio test (Muse & Weir 1992) is a versatile and powerful test used in phylogenetic analysis to test if a model is significantly better than an alternative model. Other tests can be used with the some purpose and in different circumstances, for instance the relative rate test (Sarich & Wilson 1967), but the overall performance of the likelihood ratio test seems to be similar or better than the others (Muse & Weir 1992; Tajima 1993).

The likelihood ratio test is commonly used in at least four different situations.

1 -The first use, to test incongruence between data sets
2 - to test nested models, a particular case of which is to assess whether a molecular clock hypothesis adequately describes the data
3- to test the fit of a maximum likelihood model to the observed data
4- to compare different evolutionary tree topologies.

The likelihood ratio test statistic (Huelsenbeck & Crandall 1997)

$$\Lambda = \frac{\max\left[L0\ (\text{Null Model} \mid \text{Data})\right]}{\max\left[L1\ (\text{Alternative Model} \mid \text{Data})\right]}$$

for practical reasons, the minus log likelihood is used and the statistic becomes:

$$-\log \Lambda = \log L1 - \log L0$$

The likelihood ratio test statistic (Huelsenbeck & Crandall 1997)

Where *L1* is the maximum likelihood of the alternative hypothesis, (the complex, parameter rich hypotheses) and *L0* is the maximum likelihood of the null hypothesis (the simpler hypothesis). For nested models, when one of the models is a particular case of the other, obtained by constraining one or more parameters of the alternative hypothesis, $-2\log\Lambda$ (the notation 2δ is also common) approximates to a $\chi 2$ distribution with degrees of freedom equal to the difference in the number of free parameters in the two models.
The statistic becomes:

$$-2\log \Lambda = 2\delta = 2(\log Lgeneral - \log Lnested)$$

Likelihood ratio test – molecular clock

A particular case of the nested model is the test for the molecular clock hypothesis, which is equivalent to comparing the likelihood of an additive tree with the one of a nested ultrametric tree. If the sequences were evolving at similar rates an ultrametric tree would not be significantly different from an additive tree, but if the rates were different then an additive tree would be significantly better than the ultrametric one. The statistic is again $\chi 2$ distributed with n-2 degrees of freedom, where n is the number of sequences, and it corresponds to the difference in the number of branch lengths that have to estimate in an additive and in an ultrametric tree.
The statistic becomes:

$$2\delta = 2(\log \textit{Lno clock} - \log \textit{Lclock})$$

Tree topology tests

Three main tree topology tests are used:
1.Kishino-Hasegawa(KH) test
2.Shimodaira–Hasegawa(SH) test
3.Approximately unbiased (AU) test
The **KH test** (Kishino and Hasegawa, 1989) was designed to test 2 trees and thus has no correction for multiple testing. This is solved in the **SH test** (Shimodaira and Hasegawa, 1999). However, the SH test becomes too conservative when testing many trees.
The **AU test** (Shimodaira, 2002) fixes this issue and is thus recommended as replacement for both KH and SH tests.

Kishino-Hasegawa test (Kishino and Hasegawa 1989)

Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999)

Shimodaira-Hasegawa (SH-test)

(Shimodaira and Hasegawa, 1999)

The test statistic is the score difference between the Maximum Likelihood tree and every other tree compared:

i.e., $\delta_T = lnL_{ML} - lnL_T$

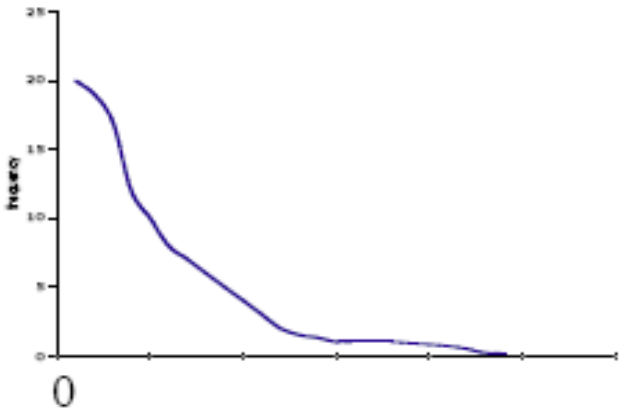
Hypotheses that we wish to test are:

H_0: all trees are equally good explanations of the data

H_1: some or all trees are not equally good explanations of the data

What is the expected distribution of δ under the null?

Hint: We know $lnL_{ML} \geq lnL_T$



Shimodaira-Hasegawa test

The SH test

Shimodaira and Hasegawa (1999) have described a resampling method that approximately corrects for testing multiple trees. They suggest that we

1. Make R bootstrap samples of the N sites. For each compute the total log-likelihood. (This is most conveniently done by RELL sampling where we add up sitewise log-likelihoods without re-estimating branch lengths or other parameters.)
2. For each tree, subtract from the sum of the resampled log-likelihoods its mean across all R bootstrap samples. This "centering" has the effect of adjusting all trees so their resampled log-likelihoods have the same expectation. Thus if the total log-likelihood of the ith tree in the jth bootstrap sample is $\tilde{\ell}_{ij}$, compute the centered value for it as

$$\tilde{R}_{ij} = \tilde{\ell}_{ij} - \frac{1}{R}\sum_{k=1}^{R}\tilde{\ell}_{ik} \qquad (21.1)$$

3. For the jth bootstrap replicate, compute for the ith tree how far that centered value is below the maximum across all trees for that replicate:

$$\tilde{S}_{ij} = \left(\max_{k}\tilde{R}_{kj}\right) - \tilde{R}_{ij} \qquad (21.2)$$

4. For each tree i, the tail probability is then taken to be the fraction of the bootstrap replicates in which $\tilde{S}_{ij}$ is less than the actual difference between the maximum likelihood and the log-likelihood L_i of that tree.

Approximately Unbiased Test

Syst. Biol. 51(3):492–508, 2002
DOI: 10.1080/10635150290069913

An Approximately Unbiased Test of Phylogenetic Tree Selection

HIDETOSHI SHIMODAIRA

Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minatoku, Tokyo 106–8569, Japan;
E-mail: shimo@ism.ac.jp

Abstract.— An approximately unbiased (AU) test that uses a newly devised multiscale bootstrap technique was developed for general hypothesis testing of regions in an attempt to reduce test bias. It was applied to maximum-likelihood tree selection for obtaining the confidence set of trees. The AU test is based on the theory of Efron et al. (Proc. Natl. Acad. Sci. USA 93:13429–13434; 1996), but the new method provides higher-order accuracy yet simpler implementation. The AU test, like the Shimodaira–Hasegawa (SH) test, adjusts the selection bias overlooked in the standard use of the bootstrap probability and Kishino–Hasegawa tests. The selection bias comes from comparing many trees at the same time and often leads to overconfidence in the wrong trees. The SH test, though safe to use, may exhibit another type of bias such that it appears conservative. Here I show that the AU test is less biased than other methods in typical cases of tree selection. These points are illustrated in a simulation study as well as in the analysis of mammalian mitochondrial protein sequences. The theoretical argument provides a simple formula that covers the bootstrap probability test, the Kishino–Hasegawa test, the AU test, and the Zharkikh–Li test. A practical suggestion is provided as to which test should be used under particular circumstances. [Approximately unbiased test; confidence limit; Kishino–Hasegawa test; maximum likelihood; multiscale bootstrap; phylogenetics; selection bias; Shimodaira–Hasegawa test.]

Combine Data

One of the most common questions in phylogenetic analysis is whether and how to combine different data sets or do a separate analysis for each one. Three general solutions have been adopted: total evidence, always combining the data sets (Kluge 1989); separate analysis, always analysing each data set and comparing the trees produced by each one, also called congruence or consensus approach; and conditional combination. The latter only combines data sets after testing them for data heterogeneity, whether differences among trees can be or not be explained by stochastic variation (Huelsenbeck *et al.* 1996a). If there is congruence between data sets then combining them would make the most use of the available information. Each of these approaches has its advantages and disadvantages, but conditional combinations are less based on philosophical assumptions and has a stronger rationality behind it.

Supertree vs Supermatrix

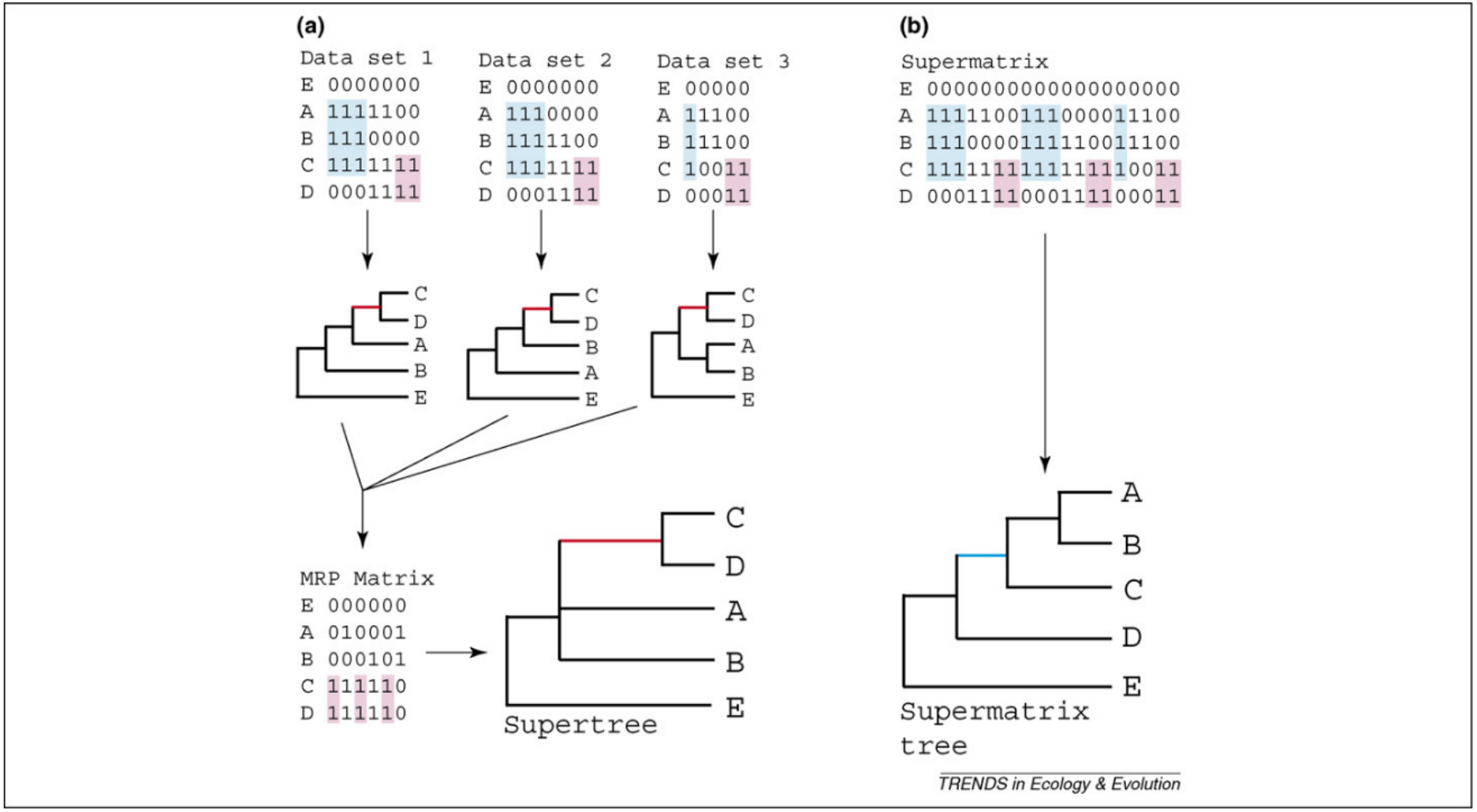


Figure 1. Schematic of MRP supertree (left) and parsimony supermatrix (right) approaches to the analysis of three data sets. Clade C+D is supported by all three separate data sets, but not by the supermatrix. Synapomorphies for clade C+D are highlighted in pink. Clade A+B+C is not supported by separate analyses of the three data sets, but is supported by the supermatrix. Synapomorphies for clade A+B+C are highlighted in blue. E is the outgroup used to root the tree.

Queiroz and Gatesy 2006 TREE

Genes rate heterogeneity

Traditionally, phylogenetic analyses over many genes combine data into a contiguous block. Under this concatenated model, all genes are assumed to evolve at the same rate. However, it is clear that genes evolve at very different rates and that accounting for this rate heterogeneity is important if we are to accurately infer phylogenies from heterogeneous multigene data sets. There remain open questions regarding how best to incorporate gene rate parameters into phylogenetic models and which properties of real data correlate with improved fit over the concatenated model. In this study, two methods of accounting for gene rate heterogeneity are compared: the n-parameter method, which allows for each of the n gene partitions to have a gene rate parameter, and the a-parameter method, which fits a distribution to the gene rates. Results demonstrate that the n-parameter method is both computationally faster and in general provides a better fit over the concatenated model than the a-parameter method. Furthermore, improved model fit over the concatenated model is highly correlated with the presence of a gene with a slow relative rate of evolution. [AIC; gene rates; phylogenetic integration; phylogenomics; rate heterogeneity]

Bevan *et al.* 2007 SysBiol

Incongruence Length Difference ILD

Two main tests have been proposed for conditional combination, the incongruence length difference (Mickevich & Farris 1981; Farris *et al.* 1995; Cunningham 1997b; 1997a) and the likelihood heterogeneity test (Huelsenbeck & Bull 1996).
The first test is based on the Mickevich-Farris index of incongruence among data sets, and the test statistic is simple:

$$I = Lc - \sum_{i=1}^{n} Li$$

Where *Lc* refers to the length of the most parsimonious tree from the combined analysis, and *Li* is the length of the most parsimonious tree on the *i*-th data set, out of a total of *n* data sets. The *I* value is then compared with the distribution of *I* values expected from chance alone.

Likelihood Heterogeneity Test

The likelihood heterogeneity test is a likelihood ratio test where the first likelihood, the alternative hypothesis, is the one of the tree when different trees can underline each data partition, and the second, the null hypothesis, is the likelihood of the tree when the same trees are assumed to underlie all data partition, in spite of possible differences in evolution rates and parameter values. The null distribution is calculated using simulation and the significance of the log likelihood assessed.
If the results of the test are not significant, meaning differences between independent data sets trees were only due by chance, then combination analysis of the data set can be carried out.