Influence Function for Robust Phylogenetic Reconstructions

Avner Bar-Hen, Mahendra Mariadassou,* Marie-Anne Poursat,† and Philippe Vandenkoornhuyse‡*

*Univ. René Descartes, MAP5, 45 rue des Saints-Pères, 75270 Paris, France; †Univ. Paris-Sud Bât 425, Dept. de maths, Orsay, Paris, France; and ‡Univ. de Rennes I, UMR 6553 Ecobio, Rennes, Paris, France

Based on the computation of the influence function, a tool to measure the impact of each piece of sampled data on the statistical inference of a parameter, we propose to analyze the support of the maximum-likelihood (ML) tree for each site. We provide a new tool for filtering data sets (nucleotides, amino acids, and others) in the context of ML phylogenetic reconstructions. Because different sites support different phylogenic topologies in different ways, outlier sites, that is, sites with a very negative influence value, are important: they can drastically change the topology resulting from the statistical inference. Therefore, these outlier sites must be clearly identified and their effects accounted for before drawing biological conclusions from the inferred tree. A matrix containing 158 fungal terminals all belonging to Chytridiomycota, Zygomycota, and Glomeromycota is analyzed. We show that removing the strongest outlier from the analysis strikingly modifies the ML topology, with a loss of as many as 20% of the internal nodes. As a result, estimating the topology on the filtered data set results in a topology with enhanced bootstrap support. From this analysis, the polyphyletic status of the fungal phyla Chytridiomycota and Zygomycota is reinforced, suggesting the necessity of revisiting the systematics of these fungal groups. We show the ability of influence function to produce new evolution hypotheses.

Introduction

Phylogenetic methods are used in many diverse fields, including molecular evolution, virology, and ecology. Maximum likelihood (ML) is one of the most popular. It is based on the adoption of an explicit DNA or protein sequence evolution model. Depending on the complexity of the model, the inferred tree can be very dependent on randomly occurring peculiarities in the data set; thus, its robustness must be assessed. The most commonly used test of reliability of an inferred tree is the bootstrap (Efron 1979; Felsenstein 1985), though the simulation output is, unfortunately, rarely examined to determine whether their conclusions are only driven by a few peculiar sites.

Empirical research in many areas of statistics gives high priority to detecting outliers. Indeed, outliers have a strong effect on the results of a statistical analysis and can even invalidate conclusions drawn from them. In molecular phylogenetics, every site takes part in the inference of a phylogenetic tree. But how stable is the inferred tree? In other words, are there any sites that drive the tree topology, thus inducing change(s) when deleted? Does the support of a branch rest on an atypical segment of the DNA sequence? Drawing valid conclusions from a phylogenetic tree requires to control these outlier sites. Although the classical emphasis is to minimize the influence of such sites, the most interesting aspect might be to *detect* them. Influence functions, introduced by Hampel (1974) as a measure of the impact that each piece of sampled data has on the statistical inference, are helpful to detect such influential segments of sequence. In this paper, we make use of the influence function concept to obtain influence diagnosis in phylogeny. Various other uses of the influence function can be found in Huber (2004), and the relationships between jackknife and influence function were proved in Miller (1974).

Resampling techniques are the most widely used approaches to assess the stability of inferred trees, but there are other approaches that have been used to assess robustness in the context of phylogenetic analyses. For example, Archibald and Roger (2002) used a likelihood ratio test for scanning DNA sequence alignments to detect regions of incongruent phylogenetic signals, such as those influenced by recombination. Blouin et al. (2005) presented a simulation study in which they evaluated the robustness of evolutionary site-rate estimates for both small and phylogenetically unbalanced samples.

Because we want to characterize the influence of each site on the likelihood, it is crucial to study them one at a time. Let T be the tree that maximizes the likelihood of the whole data set and $T^{(h)}$ be the tree that maximizes the likelihood of the jackknife sample obtained when removing site h from the original data set. By comparing T to each $T^{(h)}$, we study the impact of each site on T and can relate the stability or lack of a stability of a clade to a particular site or set of sites. We also define the outlier sites as those whose influence values are the greatest. Outlier sites may arise from biological well-known characteristics that result in evolution schemes not taken into account by the evolution model, such as the nature of mutation of GC content for a given nucleotide data set. Taking a further step toward robustness, we order the sites in the original data set from strongest outlier to weakest outlier and remove them one at a time, starting with the strongest outlier. Doing so, we obtain a sequence of samples, each one shorter than the previous one by exactly one nucleotide, from which the corresponding sequence of trees is inferred. Assuming that major causes of disruption and thus instability disappear along with the strongest outlier, we expect a stable tree to arise from this sequence. The main issue is then: how many outliers must be removed before the inferred tree becomes robust?

Materials and Methods

Definitions and Notations

Let us consider s homologous nucleotide sequences that consist of n nucleotide sites to construct a tree. Let $\mathbf{X} = (X_{pq})$ be the $s \times n$ matrix of data where X_{pq} is T, C, A, or G and denotes the state of the qth site in species

Key words: nfluence function, phylogenetic, maximum likelihood, tree stability.

E-mail: avner@math-info.univ-paris5.fr.

Mol. Biol. Evol. 25(5):869–873. 2008
doi:10.1093/molbev/msn030
Advance Access publication February 7, 2008

p. Let $\mathbf{X}_h = (X_{1h}, \ldots, X_{sh})$ be the data at the hth site. The superscript denotes the transpose operator.

Assuming a substitution model and independently evolving sites, the log-likelihood of a given tree T is

$$l_T(\theta_T|\mathbf{X}) = \sum_{h=1}^{n} \log f_T(\mathbf{X}_h|\theta_T), \quad (1)$$

where $f_T(\mathbf{X}_h|\theta_{Tgr;})$ is the probability to observe pattern, that is, alignment column, $\mathbf{X}_h$ at the homologous site h. We note that the log-likelihood divided by the sample size, $l_T(\theta_T|\mathbf{X})/n$, can be regarded as an unbiased estimator of the expected log-likelihood per site. Even if the sites are correlated, it is an unbiased estimator of the expected log-likelihood per site, under mild assumptions on the correlation structure (e.g., ergodicity of the Markov chain modeling the correlation) (Bar-Hen and Kishino 2000).

Given the topology describing the branching order, the log-likelihood is expressed in terms of the transition probabilities computed from the evolution model. The vector θ_T denotes the set of unknown parameters such as the branch lengths of tree T and the substitution rate of the evolution model. We refer to Bryant et al. (2005) for an up-to-date review on ML techniques for phylogenetics.

Influence Function for Phylogeny

We adapt the concept of influence function to the context of phylogenetics. To a given alignment $\mathbf{X}=(\mathbf{X}_h)_{h=1,\ldots,n}$, we associate the log-likelihood statistic:

$$S(F_n) = \frac{1}{n}\sum_{h=1}^{n} \log f_T(\mathbf{X}_h|\theta_T),$$

with $f_T(X|\theta_T)$ defined in equation (1) and where T is the tree maximizing the likelihood of $\mathbf{X}$.

The effect of deleting site $\mathbf{X}_h$ can be measured by its influence value $IF_{S,F_n}(\mathbf{X}_h)$:

$$IF_{S,F_n}(\mathbf{X}_h) = (n-1)(l_T(\theta_T|\mathbf{X}) - l_{T^{(h)}}(\theta_{T^{(h)}}|\mathbf{X}^{(h)})), \quad (2)$$

with $\mathbf{X}^{(h)}$ representing all the sites of $\mathbf{X}$, but $\mathbf{X}_h$ and $T^{(h)}$ defined in the same way as T as the tree maximizing the likelihood of $\mathbf{X}^{(h)}$. The value $IF_{S,F_n}(\mathbf{X}_h)$ gives the (scaled) change in average likelihood resulting from removing site $\mathbf{X}_h$. If a site has a positive value, this means that the parameters estimated on all sites, including the new one, has a higher likelihood than the parameters estimated on all sites but the new one. And the opposite, if a site has a negative value.

The most interesting property of equation (3) is the possibility to characterize the sites with a strong influence, that is, sites for which $IF_{S,F_n}(\mathbf{X}_h)$ is either very positive or very negative. A very positive influence value implies that the site strengthens the support for topology T, whereas a very negative value implies that the site weakens the support of topology T. In real case data set, and under our assumption that only a few sites disrupt the robustness of the inferred topology, we expect to find many sites with small positive influence value and a few sites with large negative influence value. Therefore, we focus on sites with very negative influence value and call them *outlier sites*.

Stability of the ML Tree among Trees Maximizing the Likelihood of Pseudosamples

The bootstrap is the most popular method in phylogenetics to assess the uncertainty of the inferred tree. Using pseudosamples, *P*values are computed for the branches of the tree. These *P*value are intended to estimate the support provided by the data to a clade. They can be used to build a majority-rule consensus tree in which only clades with a *P*value greater than 0.5 appear. The jackknife and influence function provide additional information to the stability of clades. Mainly, they relate the stability of a clade to certain particular sites. Thus, original information can be extracted. For example, do the outliers have a specific nucleotide content?

Bootstrap analysis, just like any statistical analysis, is sensitive to individual observations. In a phylogeny analysis, questions such as "would the support of that clade differ if these sites were discarded from the analysis? " or "are the clades sensitive to the considered sample? " often arise. To answer them, it is important to focus on the effect of individual sites on bootstrap values. Empirical influence values are useful in this context as they can identify influential sites (i.e.,outliers).

Relationship between Influence Function and the Jackknife

Let $X_1, \ldots, X_n$ be random variables with common distribution function (df) F on $\mathbb{R}^d$ ($d \geq 1$). To simplify notations, we use distribution function and probability measure indifferently: F is either one or the other. Suppose that we are interested in a parameter that can be expressed, as often in statistics, as a functional $S(F)$ of the generating df, S being defined on the space F of dfs. To evaluate the importance of an additional observation $x \in \mathbb{R}^d$, we can define, under conditions of existence, the quantity

$$IF_{S,F}(x) = \lim_{\in \to 0} \frac{S((1-\in)F + \in \delta_x) - S(F)}{\in}, \quad (3)$$

which measures the influence of an infinitesimal perturbation on the functional $S(F)$ along the direction δ_x (Efron 1979). δ_x is the Dirac measure that concentrates the whole probability mass 1 on the point x. The influence function $IF_{S,\ F}(x)$ is defined pointwise by equation (3), if the limit exists for every x.

Usually F is unknown, so that one has to estimate it by the empirical distribution function defined from the sample as:

$$F_n = \frac{1}{n}\sum_{h=1}^{n} \delta_{X_h}.$$

The natural estimator of $S(F)$ is then $S(F_n)$, and the empirical version of the influence function is obtained from equation (3) by replacing F with F_n. The particular values $IF_{S,F_n}(X_h)$ are called the empirical influence values. There is a strong connection between the influence function and the jackknife (Miller 1974; Efron 1979), which is a statistical technique for empirically estimating the variability of an estimator. The jackknife involves dropping one observation from the sample at a time and calculating the corresponding estimate

each time. Let $F_{n-1}^{(h)} = \frac{1}{n-1}\sum_{j,j\neq h}\delta_{X_j}$ be the empirical df calculated with X_h omitted from the data. Then, $F_n = \frac{n-1}{n}F_{n-1}^{(h)} + \frac{1}{n}\delta_{X_h}$ and a numerical approximation of $IF_{S,F_n}(X_h)$ can be obtained using $\in = \frac{-1}{(n-1)}$:

$$\begin{aligned} IF_{S,F_n}(X_h) &\approx \frac{S((1-\in)F_n + \in\delta_{X_h}) - S(F_n)}{\in} \\ &= (n-1)(S(F_n) - S(F_{n-1}^{(h)})) \\ &= S_{n,h}^* - S(F_n), \end{aligned}$$

where $S_{n,h}^* = nS(F_n) - (n-1)S(F_{n-1}^h)$ are the pseudovalues of the jackknife, that is, the estimated values of $S(F)$ computed on $n-1$ observations (Miller 1974).

An alternative to influence function to measure the impact of site $\mathbf{X}_h$ on the inference of a statistic S is *jackknife-after-bootstrap*: the value of S over the whole sample is compared with the values $S_1^*, \ldots, S_B^*$ obtained from bootstrap samples where $\mathbf{X}_h$ does not occur. However, the computational time involved in most ML techniques makes it demanding, in time and in computer resources, to perform bootstrap analyses. In addition, influence functions are anchored in a more classical framework. Therefore, we favored influence function over jackknife-after-bootstrap.

Data set

The influence function of each site was computed from an alignment of the small subunit rRNA gene (1 026 nt) over 157 terminals (i.e., 157 rows), all fungi belonging to the phyla Chytridiomycota, Zygomycota, Glomeromycota plus one outgroup to root the tree, *Corallochytrium limacisporum*, a putative choanoflagellate. This alignment, previously published in Vandenkoornhuyse et al. (2002), was chosen to satisfy different criteria: 1) enough variation accumulated to clearly resolve the phylogenetic topology, 2) a very low number of detectable homoplasic events, 3) a strong monophyletic group (i.e., Glomeromycota), 4) a highly polyphyletic group (i.e., Zygomycota), and 5) one group with uncertainties about phylogenetic affinities (i.e., Chytridiomycota).

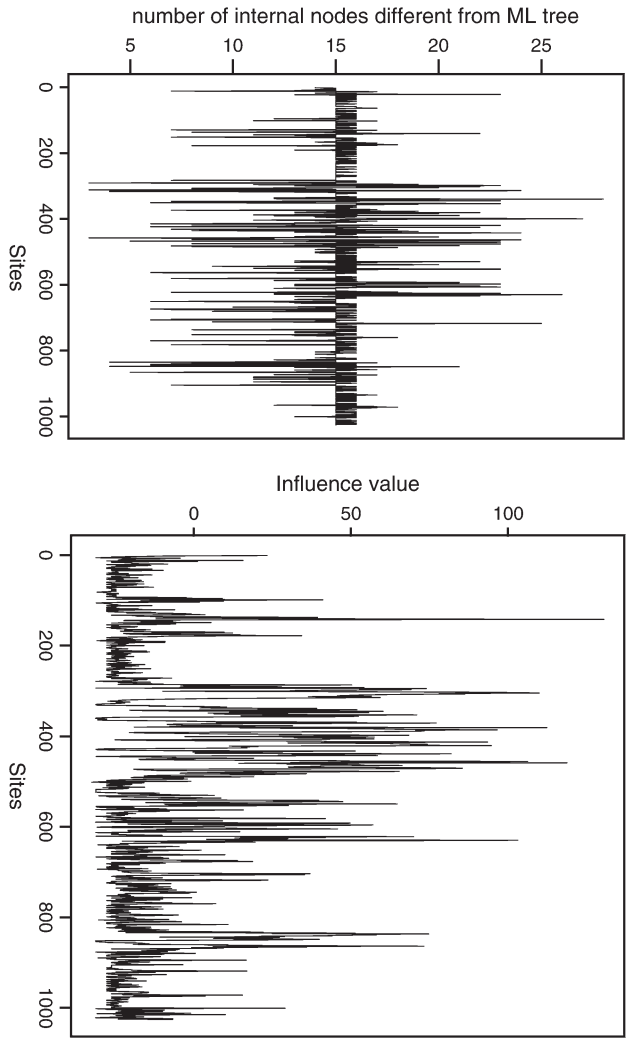

FIG. 1.—(*A*) Number of internal nodes different from the ML-GTR guide tree (all data included) when removing one site only from the data set. (*B*) Influence values when removing each of the single sites (i.e., one column only) from the data set (1 026 columns in total).

Results and Discussion

In this paper, we focused on the detection of influential sites (i.e., outliers) for the ML tree of fungi belonging to the phyla Chytridiomycota, Zygomycota, and Glomeromycota. The idea developed here is that computing influence values helps to detect outliers for the proposed model of evolution and to compute a more robust tree.

The influence function of each site was computed from an alignment containing 157 fungal terminals and 1 026-nt sites (i.e., 1 026 columns and 157 rows) (see Data set).

We first performed an ML estimation of the phylogeny of the 158 sequences using the PHYML program (Guindon and Gascuel 2003). The ML tree T was constructed with the general time reversible (GTR) model (Felsenstein 2004). Furthermore, we have evaluated the fit to our data of different models of nucleotide substitution (including HKY, F81, JC, etc.) using "modeltest" (Posada and Crandall 1998) (http://darwin.uvigo.es/software/modeltest.html) and confirmed the validity of the choice of the GTR model. The tree presented in supporting online material is in accordance with previously published trees (Vandenkoornhuyse et al. 2002) and provides a result congruent to the maximum parsimony tree.

We used an R script to compute the influence values equation (2) for each of the 1 026 sites of the alignment. (All scripts written with *R* software available upon request to the A.B-H.) Each influence value is computed by removing one site h from the whole data set, computing the ML tree $T^{(h)}$ on the obtained jackknife sample and taking the difference between the mean likelihood of a site under the ML tree T and under $T^{(h)}$. We found out that certain sites have very negative influence values, that is, removing them strongly worsens the likelihood of the ML tree (fig. 1). Furthermore, some the $T^{(h)}$ were quite different from T. In other words, when removed, some sites significantly modified the inferred tree. Figure 1 plotted, for each site h, the number of internal nodes of the ML tree T not found in tree $T^{(h)}$. This proves that a change in the likelihood of a sample reflects a change in the underlying ML topology: change of topology and change of likelihood are strongly connected.

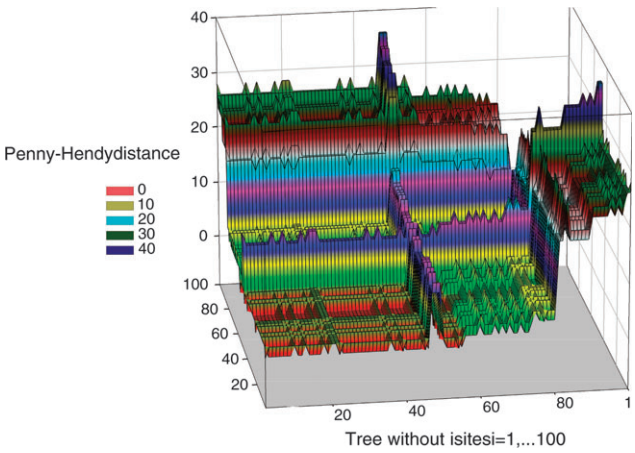

FIG. 2.—Penny–Hendy tree-to-tree distances. *x* and *y* axes figure the trees inferred after removing the *i* strongest outliers ($i = 1, \ldots, 100$). The ML-GTR guide tree (all data included, i.e., $i = 0$) is not included on this figure.

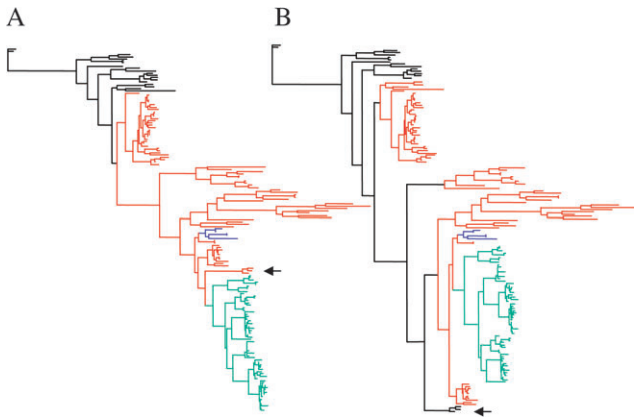

FIG. 3.—(*A*) Comparison of the tree topology for the ML-GTR guide tree versus ML-GTR minus the strongest outlier and reciprocally for (*B*) the ML-GTR minus the strongest outlier versus ML-GTR guide tree. Black, part of the tree not affected by the removal of the strongest outlier; green, phylum Glomeromycota; red, phylum Zygomycota; blue, phylum Chytridiomycota. For the Zygomycota and Chytridiomycota notice that parts remained unchanged thus colored in black. The arrow indicates the position of the 3 terminals *Mucor ramannianus*, *Umbelopsis nana*, and *Umbelopsis isabellina* as an example of a modification induced within the topology when removing the strongest outlier.

When removing a site, between 11 and 32 internal nodes of the ML tree were affected. Figure 1 showed an average of 15 nodes affected by removing only one site. These nodes were related to terminals with high homology within unresolved clades, that is, not well supported by the ML tree. Some areas contained the strongest outliers that were not uniformly distributed along the sequence.

For example, the most influential site (i.e., strongest outlier) (position 142 on the data set) corresponded to a highly variable site. To visualize the position of this particular site, we computed the most probable RNA secondary structure (RNA folding) using a method based on thermodynamic principles (Zuker et al. 1999) (mfold at http://frontend.bioinfo.rpi.edu/applications/mfold/). From 2 different sequences selected randomly, and using different temperatures and different salinities, we always found that the strongest outlier is on a small loop (5 nt) carried by a conserved hairpin (figure not shown, available on request).

In order to achieve a more robust tree, we removed the strongest outliers from the analysis. If the outliers indeed disrupt the inferred topology, we expect that, after discarding enough of them, the inferred tree will not be oversensitive to the sample anymore, that is, removing or adding one site from the analysis will not drastically change it. In order to test this belief, we classified the outliers according to their influence values, from the most negative to the least negative. We then deleted the *i* strongest outliers (for values of *i* ranging from 1 to 325) and inferred the ML-GTR tree. Using the Penny–Hendy distance (Penny and Hendy 1985), we quantified the topological similarity of these 325 trees with each other and with the guide tree *T*. Penny–Hendy distance between two phylogenies calculates the minimal composition of elementary mutations that convert the first tree into the second one. From the data set, we demonstrated that there were two stable trees. Removing any number between 2 and 44 of the strongest outliers led to almost the same tree. This is illustrated by the very small Penny–Hendy distance between these topologies (fig. 2). After removing the 46 strongest outliers, an additional stable topology was found, but the tree-to-tree distance increased quickly after removing 50 sites leading to unstable phylogenies (fig. 2).

We focus on the 325 sites with negative influence, but we can probably concentrate on fewer sites. Huber (2004) proved the asymptotic normality of the influence value under very general conditions. Using empirical mean and variance and given a type I error level, it gives a practical solution to determine the threshold.

Strikingly, removing as few as the two strongest outliers already provides an improved stability: the majority of internal nodes in common with the ML tree have better bootstrap values (results ML-GTR and K2P-NJ tree reconstruction, data not shown). This further confirms the assumption that the removed information does not contribute to the ML tree.

We take a closer look at the topologies inferred when removing the strongest outlier from the data set to understand how and where it differs from the ML tree. Figure 3 shows the high magnitude of these differences. Different interpretations transpired from the trees inferred before and after removing the strongest outliers (fig. 3). First, the phylum Glomeromycota appeared remarkably stable and monophyletic. Only slight changes in the position of terminals can be detected when the trees generated from the data set minus the strongest outlier to the data set minus the 40 strongest were compared. These changes were observed within the cluster of 13 terminals containing 3 morphological species, *Glomus mosseae*, *G. claroideum*, and *G. lamellosum*. These changes might be attributable to the fact that these terminals are closely related, and the quantity of molecular information was not high enough to clearly resolve their phylogenetic affinities. Second, the phylum Chytridiomycota appeared polyphyletic, with a group of terminals containing *Basidiobolus* (two terminals), *Neocallimastix* (four terminals), one *Spizellomyces*, one *Chytridium*, and one *Pyromyces*, which was weakly

supported by bootstrap value (i.e., 51/55, respectively, for MP and K2P-NJ) with the whole data set; but the divergence of this group from the other Chytridiomycetes was reinforced when deleting the strongest outlier (bootstrap value = 63.5% and 66%, respectively, for MP and K2P-NJ), placed among terminals of the Zygomycota group. This result indicates that systematics within Chytridiomycota and Zygomycota must be reevaluated, and this particular group must be reclassified within a Zygomycota subphylum. From these results, we argue that the 2 Chytridiomycota groups have distinct evolutionary stories.

Acknowledgments

A.B-H. and P.V. acknowledge École thématique génomique environnementale for helpful discussions. M. Bormans is acknowledged for comments on the manuscript. P.V. acknowledges Groupement d'Intérêt Scientifique "Génomique marine" and "Fondation Total" and Agence Nationale de la Recherche (ECCO: ECosphère COntinentale : processus et modélisation) for funding.

Literature Cited

Archibald JM, Roger AJ. 2002. Gene conversion and the evolution of euryarchaeal chaperonins: a maximum likelihood-based method for detecting conflicting phylogenetic signal. J Mol Evol. 55:232–245.

Bar-Hen A, Kishino H. 2000. Comparing the likelihood functions of phylogenetic trees. Ann Inst Stat Math. 42:43–56.

Blouin C, Butt D, Roger AJ. 2005. The impact of taxon sampling on the estimation of rate of evolution at sites. Mol Biol Evol. 22:784–791.

Bryant D, Galtier N, Poursat MA. 2005. Likelihood calculation in molecular phylogenetics. In: Gascuel O, editor. Mathematics of Evolution and Phylogeny. Oxford (UK): Oxford University Press. p. 33–62.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann Stat. 7:1–26.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, (MA): Sinauer Associates.

Guindon S, Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hampel FR. 1974. The influence curve and its role in robust estimation. J Am Statist Assoc. 69:383–393.

Huber PJ. 2004. Robust Statistics. Sussex (UK): Wiley.

Miller RG. 1974. The jackknife — a review. Biometrika. 61: 1–15.

Penny D, Hendy MD. 1985. The use of tree comparison metrics. Syst Zool. 34:75–82.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Vandenkoornhuyse P, Baldauf SL, Leyval C, Straczek J, Young JPW. 2002. Extensive and novel fungal diversity in plant roots. Science. 295:2051.

Zuker M, Mathews DH, Turner DH. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski J, Clark BFC, editors. RNA Biochemistry and Biotechnology. NATO ASI Series: Kluwer Academic Publishers. p. 11–43.

Amdt von Haeseler, Associate Editor

Accepted January 29, 2008