Influence function for robust phylogenetic reconstruction

Mahendra Mariadassou, Avner A. Bar-Hen

Influence Function for Robust Phylogenetic Reconstruction

Mahendra MARIADASSOU[1] and Avner BAR-HEN[2]

[1] Laboratoire MIG- INRA, Bât. 233, Domaine de Vilvert, 78352 Jouy-en-Josas, Cedex , France
mahendra.mariadassou@jouy.inra.fr
[2] Laboratoire MAP5, UMR8145 CNRS, 45 rue des saints-pères, 75270 Paris, Cedex 06, France
avner.bar-hen@mi.parisdescartes.fr

Abstract *Phylogenies in short, are the most convenient way to describe the relationship between different species and are widely used in several fields of biology: comparative genomics, epidemiology, conservation biology, etc. However, most inferences drawn from phylogenies are accurate only if the reconstructed phylogeny itself is accurate. For a given reconstruction bias, robust phylogenies are preferred to non robust ones. We are concerned here with the loss of robustness induced by outliers. One way to mitigate this loss is to detect and remove outliers from the dataset.*

We advocate the use of empirical influence functions to detect influent characters and taxa, which are prone to be outliers, and their removal from the data set to build robust phylogenies. Three data sets (Zygomycetes, placental mammals, T-box gene family) show that maximum likelihood phylogenies are not robust and that removing as few as a handful of outliers can significantly increase the robustness of a tree, as measured by average bootstrap values.

Keywords Biostatistics, Phylogeny, Influence Function, Outliers, Robustness.

1 Introduction

Phylogenies are an essential tool in many fields of biology and it is thus crucial to reconstruct accurate phylogenies and moreover to assess the uncertainty associated with these phylogenies. The most frequent way of doing so is to use bootstrap replicates of the alignment and to compute bootstrap values [1] of inner branches. This approach produces a global index of uncertainty that captures, among others, the variability induced by sampling of characters. However, bootstrap probabilities should be handled with caution as they do not have a clear-cut statistical interpretation [2]. Moreover, the sampling of character is not the only cause for uncertainty in the inferred phylogeny: taxon sampling is also known to impact the accuracy of phylogenetic analysis[3]. Outlying characters resulting for example from alignment artifacts as well as rogue taxa can introduce bias in the reconstruction process which leads. If the bias is strong enough, measures of variability based on random resampling, such as bootstrap values, can be blind to the influence of these characters. Here, we use influence function to systematically investigate the influence of a given character and/or a given taxon on the inferred phylogeny.

2 Methods

We work with the maximum likelihood (ML) framework under which all characters $\mathbf{X} = (X_1, \ldots, X_n)$ of an alignment are considered as random variables independently drawn from the same distribution Q on a sample space $\mathcal{A}$ (for an alignment made of s taxa and nucleotide characters, $\mathcal{A} = \{A, C, G, T\}^s$). To each topology T with branch lengths $\mathbf{b}_T$, we can associate a probability distribution $P(.; T, \mathbf{b}_T)$ on $\mathcal{A}$. The goal of ML phylogenetic inference is to find the tree $(\hat{T}, \hat{\mathbf{b}}_{\hat{T}})$ that minimizes the Kullback-Leibler divergence between Q (unknown and replaced by the empirical distribution Q_n of the X_i) and $P(.; T, \mathbf{b}_T)$ or similarly that maximizes the per-character log-likelihood $L(X_1, \ldots, X_n; T, \mathbf{b}_T)$ of the alignment under tree $(T, \mathbf{b}_T)$:

$$L(\mathbf{X}; T, \mathbf{b}_T) = L(X_1, \ldots, X_n; T, \mathbf{b}_T) = \frac{1}{n} \sum_{i=1}^{n} \log P(X_i; T, \mathbf{b}_T).$$

Using definitions first introduced in the robustness literature [4], we define the influence $IF(X_i)$ of character X_i as the normalized shift in per-character log-likelihood induced by the removal of that character:

$$IF(X_i) = (n-1)[L(\mathbf{X}; \hat{T}, \hat{\mathbf{b}}_{\hat{T}}) - L(\mathbf{X}_{-i}; \hat{T}_{-i}, \hat{\mathbf{b}}_{\hat{T}_{-i}})]$$

where $\mathbf{X}_{-i}$ is the alignment deprived of character X_i and $(\hat{T}_{-i}, \hat{\mathbf{b}}_{\hat{T}_{-i}})$ is the tree reconstructed from $\mathbf{X}_{-i}$. Characters with a high positive $IF(X_i)$ have a phylogenetic signal that strongly conflicts the signal coming from the rest of the alignment and are potential outliers.

Similarly, we define the influence $TII(T_j)$ of taxon T_j as

$$TII(T_j) = d(\hat{T}^{-j}, \widehat{T^{-j}})$$

where d is a distance between trees, $\hat{T}^{-j}$ is the tree reconstructed on the complete alignment and then pruned of taxon T_j and finally $\widehat{T^{-j}}$ is the tree reconstructed on the alignment deprived of taxon T_j from the start. Taxa with a high $TII(T_j)$ strongly change the topology when included in the alignment and are potential rogue taxa.

3 Results

We applied our method to three datasets to detect outliers and propose alternative phylogenies: 16S rRNA from fungi (Zygomycetes and Chytridiomycetes), mtDNA from placental mammals and the T-Box transcription factor gene family in bilaterians. Our results on Zygomycetes [5] show that outliers have a strong effect on the inferred phylogeny. The two most influential characters affect the topology in no less than 20 inner branches (out of 155) and reduce the log-likelihood of the ML tree by more than 100 units. Excluding these two characters leads to a robust topology, which has higher bootstrap values and reduced influence values for the remaining characters. Our results on placental mammals [6] show that rogue taxa also have a strong impact on the resulting topology and confirms the status of the guinea-pig as a rogue taxa for this dataset. Our results on the T-Box gene family enable us to identify a subset of taxa for which the phylogeny can be reconstructed with greater confidence (higher bootstrap values) for both recent and old branches than in the initial alignment. The resulting phylogeny is then used to ascertain the position of a new T-Box gene within the family.

References

[1] J. Felsenstein, Confidence Limits on Phylogenies: An Approach Using the Bootstrap, *Evolution*, 39:783-791, 1985.

[2] E. Susko, Bootstrap support is not first-order correct, *Systematic Biology*, 58, 211–233, 2009.

[3] T. A. Heath, S. M. Hedtke and D. M. Hillis, Taxon sampling, the accuracy of phylogenetic analyses, *J. Mol. Evol.*, 46:239-257, 2008.

[4] F. R. Hampel, The influence curve and its role in robust estimation, *JASA*, 69:383–393, 1974.

[5] A. Bar-Hen, M. Mariadassou, M.-A. Poursat and P. Vandenkoornhuyse, Influence function for robust phylogenetic reconstructions, *Molecular Biology and Evolution*, 25:869–873, 2008

[6] M. Mariadassou, A. Bar-Hen and H. Kishino, Taxon Influence Index, *Systematic Biology*, in press, 2011.