The Journal of Open Source Software

rGUIDANCE – alignment confidence score computation in R

Franz-Sebastian Krah[1, 2] **and Christoph Heibl**[2]

1 Plant Biodiversity Research Group, Department of Ecology & Ecosystem Management, Technical University of Munich, 85354 Freising, Germany **2** Bavarian Forest National Park, 94481 Grafenau, Germany

DOI: 10.21105/joss.01350

Software

- Review
- Repository
- Archive

Submitted: 07 March 2019
Published: 30 April 2019

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

R has become the toolbox of many researchers in ecology and evolutionary biology, both fields increasingly integrating molecular phylogenetic information. This modular toolbox allows to implement pipelines from sequence retrieval and alignment, to phylogeny estimation and high-end visualization of results. One essential step within this pipeline is currently missing from the **R** package ecosystem: detection of unreliably aligned regions within multiple sequence alignment (MSA). Although alignments are fundamental to phylogenetically informed analyses, they often contain extended unreliable regions. The alignment confidence score GUIDANCE demonstrated high accuracy in detecting such regions of low quality. Here we introduce the **R** package `rGUIDANCE`, which fully implements the alignment confidence algorithms GUIDANCE, HoT and GUIDANCE2. We will demonstrate the core functionality of `rGUIDANCE` and how it can be easily integrated into a phylogeny inference pipeline. `rGUIDANCE` is a free and open-source **R** package, available via GitHub at https://github.com/FranzKrah/rGUIDANCE//

Introduction

The free software environment for statistical computing and graphics, **R**, has developed as an indispensable toolbox in ecology and evolution and in data science in general (Muenchen, 2019). A myriad of introductory books to statistics, ecological and evolutionary analysis are based on **R** (e.g., https://www.r-project.org/doc/bib/R-books.html, UseR! series). Within ecology and evolution, a growing body of research relies on phylogenetic information (Cavender-Bares, Kozak, Fine, & Kembel, 2009), e.g., phylogenetic diversity or ancestral character reconstruction. Once users have gained basic knowledge, the modular **R** toolbox allows to code pipelines, e.g., for phylogeny inference, meeting individual's needs. Such a pipeline may contain the following steps and **R** packages (only some are listed): (1) sequence retrieval, either manual, semi-automated (*ape*) or automated (*phylotaR*), (2) sequence handling (*ape*, *seqinr*), (3) multiple sequence alignments (MSA) via interface functions to MSA programs (*ape*, *ips*) (4) phylogeny estimation based on MSA via interface functions to tree inference programs (*ape*, *phyclust*, *ips*) (5) divergence time estimation (*ape*); (6) comparative phylogenetic methods such as ancestral character estimation (*ape*, *phytools*, *geiger*) or phylogenetic community ecology such as phylogenetic diversity (*picante*, *vegan*, *MicEco*) and (7) phylogeny visualization (*ape*, *phytools* or *ggtree*).

Although it is possible to implement such a pipeline in the **R** framework, an essential step is currently not implemented: MSA confidence estimation, which would be situated

between step (3) and (4). Currently a typical workflow as outlined above includes a single MSA, which is assumed to be correct. However, benchmark studies show that alignment accuracy is often low, depending on the MSA program and settings used (Nuin, Wang, & Tillier, 2006; Thompson, Linard, Lecompte, & Poch, 2011). Numerous programs have been developed to filter MSA (e.g., trimAI, GBlocks) from ambiguously aligned regions, however simple filtering of alignments was shown to worsen phylogeny estimation in many cases (Tan et al., 2015). One promising option is to include alignment uncertainty in the phylogeny inference to down-weight sites of low reliability. Such an option is available via a combination of GUIDANCE (Penn, Privman, Landan, Graur, & Pupko, 2010) and RAxML (Stamatakis, 2014). The GUIDANCE column score can be passed as individual weights to each column of the alignment to RAxML via the –a flag (for a practical example: Krah et al. (2018)). However, GUIDANCE is currently not available in **R** and thus integration of alignment reliability is hampered. Here we thus introduce the **R** package `rGUIDANCE` and its core functionality. `rGUIDANCE` fully implements GUIDANCE, HoT and GUIDANCE2, which were shown to detect unreliable areas with high accuracy (Landan & Graur, 2008; Penn et al., 2010; Sela, Ashkenazy, Katoh, & Pupko, 2015). `rGUIDANCE` further implements basic MSA comparison tools using **Rcpp** to facilitate further development of the **R** toolbox in ecology and evolution.

Core package functionality

Here, we present the core functions of the `rGUIDANCE` package. `rGUIDANCE` makes use of several **R** packages that allow interaction with sequence data, e.g., *ape*, *ips*, *adephylo*. Further, `rGUIDANCE` uses **Rcpp** to implement MSA comparison tools, which are the basis for GUIDANCE computations. This includes the Cmatrix computation as well as the sum-of-pairs score (SP) (Penn et al., 2010). At the core of `rGUIDANCE` are the functions `guidance`, `HoT` and `guidance2`, which implement algorithms of the same name. Please note that the accuracy of GUIDANCE and `rGUIDANCE` scores in identifying alignment errors is almost identical (Fig. 1). Here we present how `rGUIDANCE` fits into the **R** ecosystem to build a basic pipeline for phylogeny inference integrating GUIDANCE column confidence scores (Fig. 2).

```
####################### Pipeline ######################
## [1] Use phylotaR to retrieve sequence clusters
# for detailed code see Vignette of R package

## [2] Load sequences for the first sequence cluster
cl0 <- read.FASTA("PATHTO/cl0.fas")

## [3] Use GUIDANCE to calculate column score (CS)
g <- guidance(cl0, ncore = 1, msa.exec = "/usr/local/bin/mafft")
msa0 <- g@msa
sc <- scores(g, "column_raxml", na.rm = FALSE)

## [4] Use alignment and CS for phylogeny inference
tr.w <- raxml(msa0, m = "GTRGAMMA", f = "a",
              N = 100, p = 1234, x = 1234,
              exec = "PATHTO/raxmlHPC-PTHREADS-AVX",
              threads = 12, weights = sc$column_raxml,
              outgroup = "Helvella_aestivalis")

## [5] Divergence time estimation
tr.w.ult <- chronos(tr.w$bipartitions)
```

```
## [6] Phylogeny visualization
plot(tr.w.ult)
```

For more detailed and up-to-date examples and tutorials, see the `rGUIDANCE` GitHub page and vignettes therein.

Conclusions

With the **R** package `rGUIDANCE` we hope to enrich the **R** toolbox for ecological and evolutionary research. `rGUIDANCE` provides implementations of well-performing MSA reliability score programs, GUDIANCE, HoT and GUIDANCE2. Alignment column scores can be computed and easily integrated into a phylogeny pipeline as was exemplarily demonstrated. The **R** package further provides further functions such as sum-of-pairs scores from MSA comparisons. These functions facilitate the modular development of further MSA reliability scores. Finally, we hope that more phylogeny-based analyses will integrate alignment uncertainty in the phylogeny inference step and thus decrease bias in ecological and evolutionary studies.

References

Cavender-Bares, J., Kozak, K. H., Fine, P. V., & Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*, 693–715. doi:10.1111/j.1461-0248.2009.01314.x

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874. doi:10.1016/j.patrec.2005.10.010

Krah, F.-S., Bässler, C., Heibl, C., Soghigian, J., Schaefer, H., & Hibbett, D. S. (2018). Evolutionary dynamics of host specialization in wood-decay fungi. *BMC Evolutionary Biology*, *18*(1), 119–132. doi:10.1186/s12862-018-1229-7

Landan, G., & Graur, D. (2008). Local reliability measures from sets of co-optimal multiple sequence alignments. In *Pacific symposium on biocomputing* (Vol. 13, pp. 15–24). doi:10.1142/9789812776136_0003

Muenchen, R. A. (2019). The popularity of data analysis software. Retrieved from Retrieved from: http://r4stats.com/articles/popularity

Nuin, P. A., Wang, Z., & Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, *7*(1), 471. doi:10.1186/1471-2105-7-471

Penn, O., Privman, E., Landan, G., Graur, D., & Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution*, *27*(8), 1759–1767. doi:10.1093/molbev/msq066

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic acids research*, *43*(W1), W7–W14. doi:10.1093/nar/gkv318

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*, 3940–3941. doi:10.1093/bioinformatics/bti623

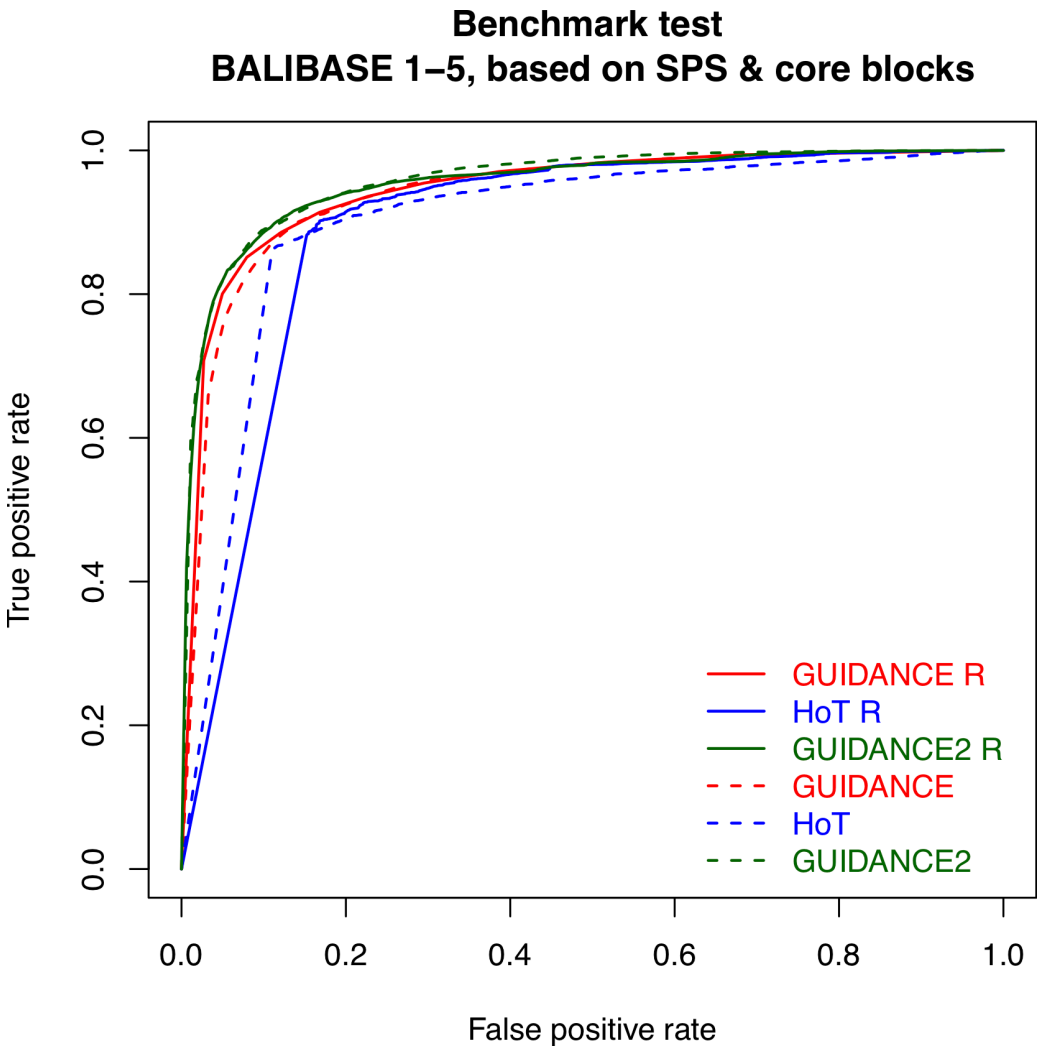

Figure 1: Accuracy of GUIDANCE scores computed with the original GUIDANCE implementation and rGUIDANCE, demonstrating equally high accuracy. Receiver operating characteristic (ROC) curves for GUIDANCE scores (red), HoT scores (blue) and GUIDANCE2 (green) of aligned residue pairs relative to the BAliBASE benchmark database (Thompson et al. (2011)). For detailed explanation of ROC curves and the BALIBASE benchmark dataset, see (Penn et al. (2010)). We followed the methods described therein. In short: We applied both implementations to each BAliBASE data set (128 datasets), using the MAFFT alignment program, generating GUIDANCE residue pair scores for each pair of aligned residues in the base MSA. We then used the BAliBASE reference MSAs in order to assess the predictive power of the residue pair scores to identify alignment errors. Each aligned residue pair in the base MSA was marked as correct/incorrect by comparing it with the reference MSA (BALiBASE). A receiver operating characteristic (ROC) analysis (Fawcett (2006)) was then applied (R package ROCR Sing, Sander, Beerenwinkel, & Lengauer (2005)) to evaluate the accuracy of the GUIDANCE confidence measure.

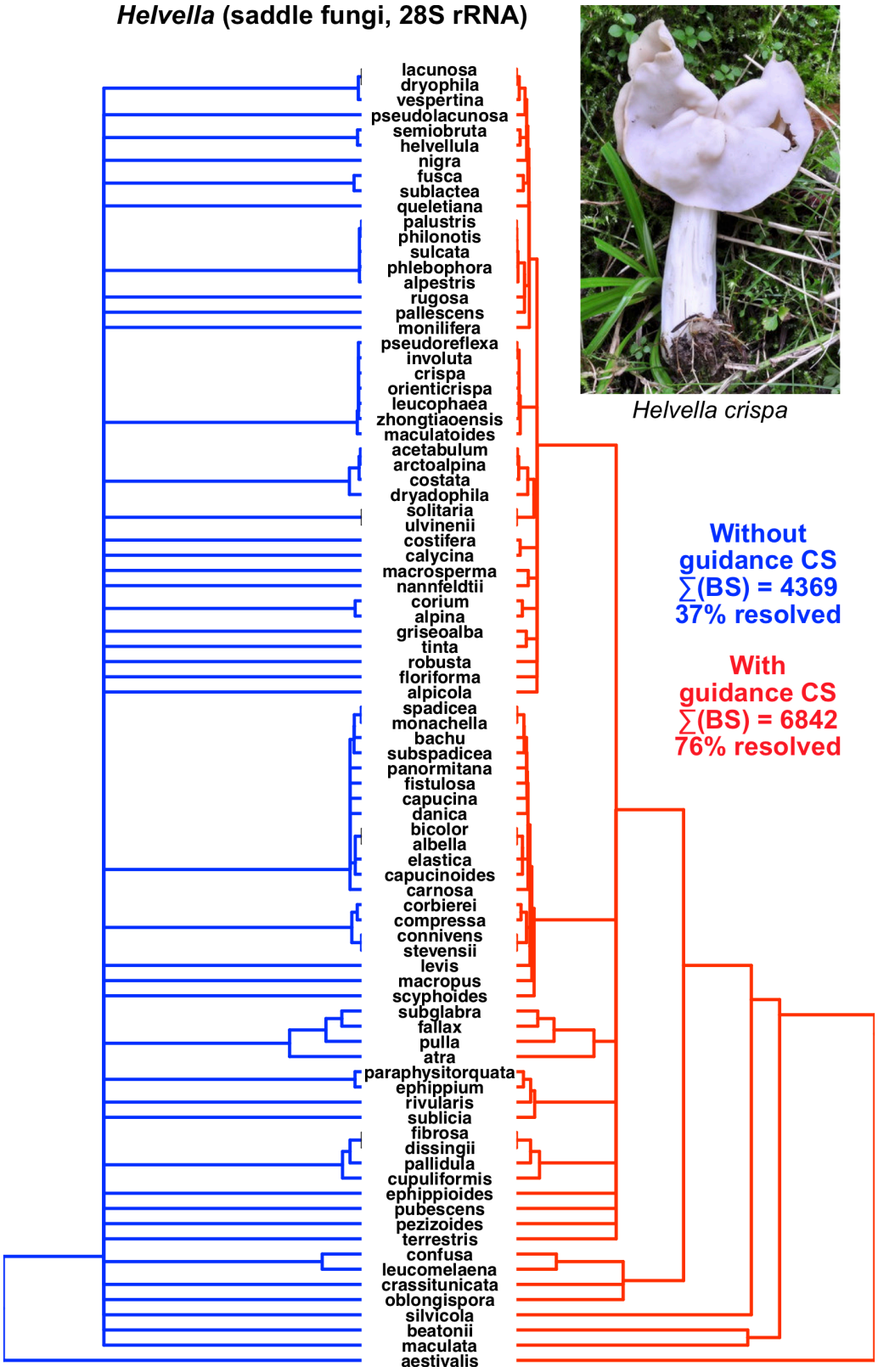

Figure 2: The estimated 28S gene tree of Helvella received substantially greater overall bootstrap support and has 39% more internal nodes resolved with a bootstrap of 70 or higher, when alignment uncertainty as represented by the GUIDANCE column score is taken into account (red), compared to the estimated topology when alignment uncertainty is ignored (blue). Photo by F.-S. Krah.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, *64*(5), 778–791. doi:10.1093/sysbio/syv033

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS ONE*, *6*(3), e18093. doi:10.1371/journal.pone.0018093