The Journal of Open Source Software

Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments

Marek L. Borowiec[1]

1 Department of Entomology, Plant Pathology and Nematology, University of Idaho

DOI: 10.21105/joss.01635

Software

- Review
- Repository
- Archive

Submitted: 31 July 2019
Published: 08 October 2019

Summary

`spruceup` is an open-source Python software package that utilizes a novel algorithm and allows flexible identification, visualization, and removal of abnormal sequences from multiple sequence alignments (MSAs). The `spruceup` algorithm aims to solve the problem of individual poorly aligned sequences rarely addressed by other alignment trimming tools. It relies on computing distance metrics in a sliding window along an MSA and comparing them to metrics across all windows for each sample. While the program is designed with trimming phylogenetic datasets in mind, its output can be used for any task requiring computing of distances along MSAs with custom parameters.

Evolutionary inference is increasingly reliant on large MSAs consisting of millions of sites from hundreds or thousands of samples. MSAs used for inference of evolutionary relationships should contain homologous nucleotides or amino acids (alignment columns) from different samples (alignment rows). However, MSAs may contain errors that violate this assumption of homology, resulting from cross-contamination, hidden paralogy, high divergence leading to insufficient information for accurate alignment, or other failures of alignment algorithms. These errors have been well-documented and shown to negatively impact phylogenetic inference. While many tools exist for trimming of poorly-aligned MSA columns, few allow identification of individual misaligned sequences. Tools for removal of poorly aligned columns address issues impacting most samples in the MSA but are not designed to identify misalignment limited to a single or few samples. Failure to remove such individual misaligned samples introduces errors into downstream analyses. Lack of tools for automated detection of these sequences means that most researchers simply do not address them and only few attempt manual alignment curation. Although visual inspection of data is always desirable, manual error correction is unsustainable with the alignment sizes used in modern phylogenetic studies. `spruceup` is intended for any phylogeneticist working with large datasets as a complement to commonly used tools that filter poorly-aligned MSA sites.

Overview

Evolutionary research relies on multiple sequence alignments (MSAs) which are matrix representations of hypotheses about the evolutionary history of genetic material. Rows in these matrices correspond to different samples, most often individual organisms representative of species or populations. The columns, which are called sites, contain nucleotides (in DNA and RNA sequences) or amino acids (in protein sequences) that should be homologous, meaning they are derived from the same ancestral nucleotide or amino acid in all included samples. Due to complexities of molecular evolution, aligning sequences such that they are homologous

is a significant computational challenge (Morrison, 2018). As a result, homology errors that have negative impact on downstream analyses, including phylogenetics and evolutionary inference (Ogden & Rosenberg, 2006), are widespread in currently published phylogenetic data (Philippe et al., 2017; Springer & Gatesy, 2018). Recent studies show that dramatically different phylogenetic trees can be produced by altering a few regions or even sites in very large MSAs (J. M. Brown & Thomson, 2017; Shen, Hittinger, & Rokas, 2017; Walker, Brown, & Smith, 2018). Because of this, researchers often post-process MSAs following alignment to identify and filter out non-homologous matrix cells.

Many tools exist that trim alignments based on low-quality sites (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009; Criscuolo & Gribaldo, 2010; Dress et al., 2008; Kück et al., 2010; Penn et al., 2010; Talavera & Castresana, 2007; M. Wu, Chatterji, & Eisen, 2012). These methods process MSAs on a column-by-column basis and can identify alignment blocks where many samples are problematic but may not have the power to identify errors affecting only one or few of the samples.

Relatively few existing alignment trimming strategies have the ability to process data on a row-by-row basis and detect errors resulting from misalignment of one or few samples. An approach that has been successfully used in several phylogenomic studies uses information from phylogenetic trees inferred from individual locus or gene alignments to identify sequences characterized by unusually long branches (Borowiec, Lee, Chiu, & Plachetzki, 2015; Yang & Smith, 2014). Unfortunately, this approach may disproportionately affect genuinely divergent sequences, such as distantly related outgroup taxa. A more sophisticated version of this idea, capable of taking into account divergent samples, is implemented in TreeShrink (Mai & Mirarab, 2018). All tree-based approaches require phylogenetic analysis prior to and following alignment trimming, which may be computationally expensive.

Two recently developed methods aim to identify artifacts affecting only one or few samples directly from sequence alignments: HmmCleaner (Di Franco, Poujol, Baurain, & Philippe, 2019) and PREQUAL (Whelan, Irisarri, & Burki, 2018). Both tools build hidden Markov models of sequences within an alignment to identify poorly-aligned sequence fragments. They are designed to work with amino acid and protein-coding nucleotide sequences and are thus not suitable for some phylogenomic alignments, such as ones constructed from partially or largely non-coding ultraconserved elements (Faircloth et al., 2012).

Finally, visual inspection of alignment followed by manual correction has been proposed to deal with alignment issues (Springer & Gatesy, 2018). Visual inspection of data is always desirable but it is unlikely to be used as a comprehensive error-removal strategy with alignments that now routinely include thousands of loci sequenced from dozens or even hundreds of samples (Edwards et al., 2016).

The Python software package `spruceup` introduced here was designed to perform fast detection of outlier sequences and alignment trimming without prior phylogenetic analysis and without penalizing genuinely divergent sequences. It is also capable of fine-scale alignment inspection that allows identification of errors affecting only parts of a locus or gene region. It fills a need for a user-friendly toolbox for flexible identification, visualization, and removal of individual misaligned sequence fragments from large multiple sequence alignments.

The workflow of outlier detection and removal is as follows:

1. Split large phylogenomic alignment into windows of user-defined size and overlap.
2. Calculate genetic distances among samples in each window; distances can optionally be scaled by distance derived from a guide tree.
3. Compute distribution of distances across windows for each sample.
4. Fit a lognormal curve to distances observed for each sample (optional).

5. Using a user-defined cutoff, reject windows from the tail of the distribution in each sample. The cutoff can represent quantile of the fitted distribution or the mean of each sample's distances across windows. The cutoff defined this way varies across samples and ensures that genuinely divergent samples are not disproportionately affected (Figure 1).

6. Plot and visualize distribution of distances and cutoffs for each sample.

`spruceup` allows the user to calculate distances across alignment windows of arbitrary size and overlap. The resulting distances are written to a `json` format file. This distance output file can be later used for exploration of various outlier cutoff criteria and thresholds, or other applications where the user needs to know sample-specific distances across the alignment. Given user-defined criteria, the removal procedure identifies outliers and records their alignment coordinates to a file. Distribution of distances across all alignment windows are visualized for each sample along with thresholds chosen for trimming. The program can be used for outlier removal with variable amount of supervision, from trimming relying on default settings to assisted manual inspection of potentially erroneous regions.

Availability and Implementation

`spruceup` is written in Python 3 and supported on Linux operating systems. It is distributed under GNU GPLv3 license (https://www.gnu.org/licenses/gpl-3.0.en.html). The source code, manual, tutorial, and example files are available on GitHub (https://github.com/marekborowiec/spruceup). `spruceup` is available on and can also be installed through the Python Package Index at (https://pypi.org/project/spruceup/).

The Journal of Open Source Software

Figure

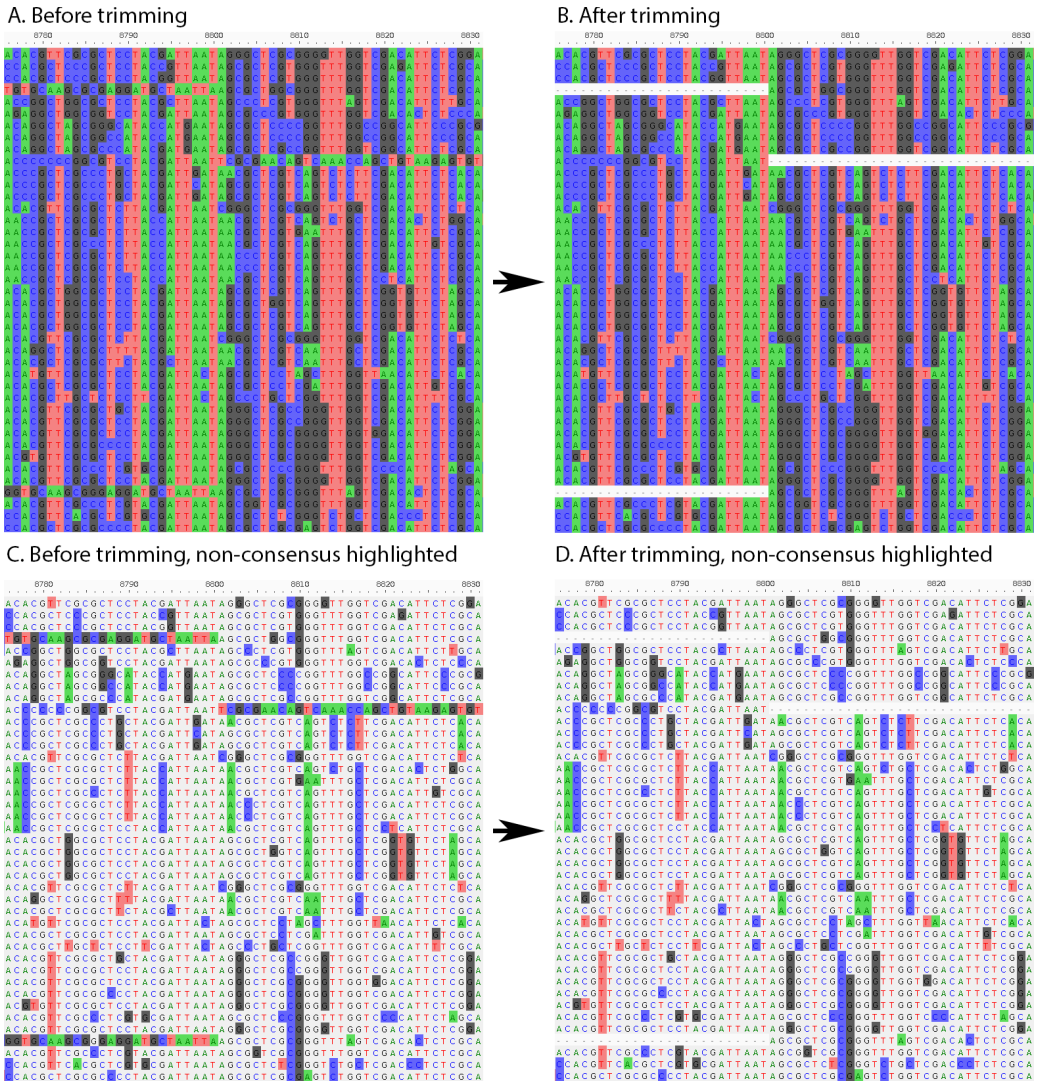

Figure 1: Outlier removal from multiple sequence alignment by `spruceup`. This nucleotide alignment was simulated under a phylogenetic tree with topology and branch lengths taken from the empirical dataset of Ultraconserved Elements (Faircloth et al., 2012) sequenced for a group of ants (Blaimer, LaPolla, Branstetter, Lloyd, & Brady, 2016). 100 alignments of 500 nucleotides each were simulated under the general time-reversible model (GTR) with state change rates between 0.1 and 1.0, base frequencies ranging from 0.1 to 0.8, and branch length scaling between 0.1 and 50, each parameter selected randomly for each alignment. After the 100 alignments were concatenated, divergent (complemented) sequence fragments were introduced (left panels), and later trimmed using the `spruceup` workflow (right panels). The figure shows only a small window of the simulated concatenated alignment of 50,000 nucleotides. Alignments visualized with AliView: https://ormbunkar.se/aliview/.

Acknowledgements

I would like to thank Philip S. Ward (UC Davis) for testing and valuable feedback on `spruceup`.

References

Blaimer, B. B., LaPolla, J. S., Branstetter, M. G., Lloyd, M. W., & Brady, S. G. (2016). Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus *Acropyga*. *Molecular Phylogenetics and Evolution*, *102*, 20–29. doi:10.1016/j.ympev.2016.05.030

Borowiec, M. L., Lee, E. K., Chiu, J. C., & Plachetzki, D. C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics*, *16*(1). doi:10.1186/s12864-015-2146-4

Brown, J. M., & Thomson, R. C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, *66*(4), 517–530. doi:10.1093/sysbio/syw101

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. doi:10.1093/bioinformatics/btp348

Criscuolo, A., & Gribaldo, S. (2010). BMGE (block mapping and gathering with entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, *10*(1), 210. doi:10.1186/1471-2148-10-210

Di Franco, A., Poujol, R., Baurain, D., & Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, *19*(1), 21. doi:10.1186/s12862-019-1350-2

Dress, A. W., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., & Stadler, P. F. (2008). Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology*, *3*(1), 7. doi:10.1186/1748-7188-3-7

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., et al. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462. doi:10.1016/j.ympev.2015.10.027

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. doi:10.1093/sysbio/sys004

Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B. M., Wägele, J. W., & Misof, B. (2010). Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology*, *7*(1), 10. doi:10.1186/1742-9994-7-10

Mai, U., & Mirarab, S. (2018). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*(S5). doi:10.1186/s12864-018-4620-2

Morrison, D. A. (2018). Multiple sequence alignment is not a solved problem. *arXiv*. Retrieved from https://arxiv.org/abs/1808.07717

Ogden, T. H., & Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, *55*(2), 314–328. doi:10.1080/10635150500541730

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., & Pupko, T. (2010). GUIDANCE: A web server for assessing alignment confidence scores. *Nucleic Acids Research*, *38*(Web Server), W23–W28. doi:10.1093/nar/gkq443

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, *283*, 1–25. doi:10.5852/ejt.2017.283

Shen, X.-X., Hittinger, C. T., & Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution*, *1*(5), 0126. doi:10.1038/s41559-017-0126

Springer, M. S., & Gatesy, J. (2018). On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity*, *16*(3), 210–228. doi:10.1080/14772000.2017.1401016

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. doi:10.1080/10635150701472164

Walker, J. F., Brown, J. W., & Smith, S. A. (2018). Analyzing contentious relationships and outlier genes in phylogenomics. *Systematic Biology*, *67*(5), 916–924. doi:10.1093/sysbio/syy043

Whelan, S., Irisarri, I., & Burki, F. (2018). PREQUAL: Detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*, *34*(22), 3929–3930. doi:10.1093/bioinformatics/bty448

Wu, M., Chatterji, S., & Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PLoS ONE*, *7*(1), e30288. doi:10.1371/journal.pone.0030288

Yang, Y., & Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, *31*(11), 3081–3092. doi:10.1093/molbev/msu245