TreeShrink: A Detector Tool for Rapid and Note Accuracy in Detecting Long Outlier Branches of Phylogenetic Trees

Mohanad A. Elagan

STEM High School for Boys - 6th of October, Egypt; mohanad.elagan1@gmail.com

Mentor: Dr. Uyen Mai - Princeton University; um6916@cs.princeton.edu

Abstract

Errors in phylogenetic tree reconstruction are found at the sequence level, where if they are overlooked, they show up as long branches in the reconstructed phylogeny. As a result, an automatic way to identify these inaccuracies is suggested in this study. It involves creating a phylogeny that can artificially identify sequences that increase the tree's diameter. Such an artificially efficient phylogeny was built using R, Python, and other programming languages. In order to locate k leaves that may be removed in order to lower the diameter of the tree, an optimization problem known as k-shrink was created using the Dendropy and BMS packages. This approach is known as TreeShrink. The statistical analyses to identify outlier species that significantly influence the tree diameter were conducted using the BMS package. This algorithm was evaluated on six phylogenomic biological datasets and an HIV dataset (strain separator), this algorithm was evaluated, and it was successful in identifying and detecting the long branches. Once the quantity of filtering is under control, TreeShrink eliminates sequences more cautiously than rogue taxon removal and frequently minimizes gene tree discordance.

Keywords: K-shrink, sequence level, rogue taxon removal

1 Introduction

Manually analyzing each sequence alignment and gene tree in detail is not practicable due to a large number of loci involved and the size of the trees. Even if possible, such manual curation presents difficulties for repeatability and is vulnerable to the biases of the curator. Sequencing is usually followed by contamination removal, homology and orthology identification, multiple sequence alignment, gene tree inference, and species tree reconstruction in phylogenetic analyses. Each phase is prone to error, and it has long been known that mistakes can spread across these processes [4][9][14].

Rogue taxon removal (RTR) and gene tree filtering are two frequently used techniques for filtering based on phylogenetic trees [10][1][16][24][7]. Filtering specific sites with a significant influence on the tree topology is one of the more recent methods [26]. By evaluating the stability concerning replication trees produced by bootstrapping [1] [16], RTR seeks to identify species that have an unstable location in the inferred trees. A second approach is to exclude genes that are thought to be troublesome, either as a result of missing data [8] [22] or signal inconsistencies [26].

In rare circumstances, the branch lengths of an inferred phylogeny might provide signs of a problem with the sequencing data. Many different sorts of phylogenetic data errors, such as incorrect orthologs, might result in the insertion of excessively long branches into the tree (for instance, figure 1a). However, rooting can be complex and error-prone [12].

The k-shrink problem: for every $1 \leq i \leq k$, identify a set of i leaves that should be eliminated in order to minimize the diameter of the tree given a tree with n leaves and a number $1 \leq k \leq n$.

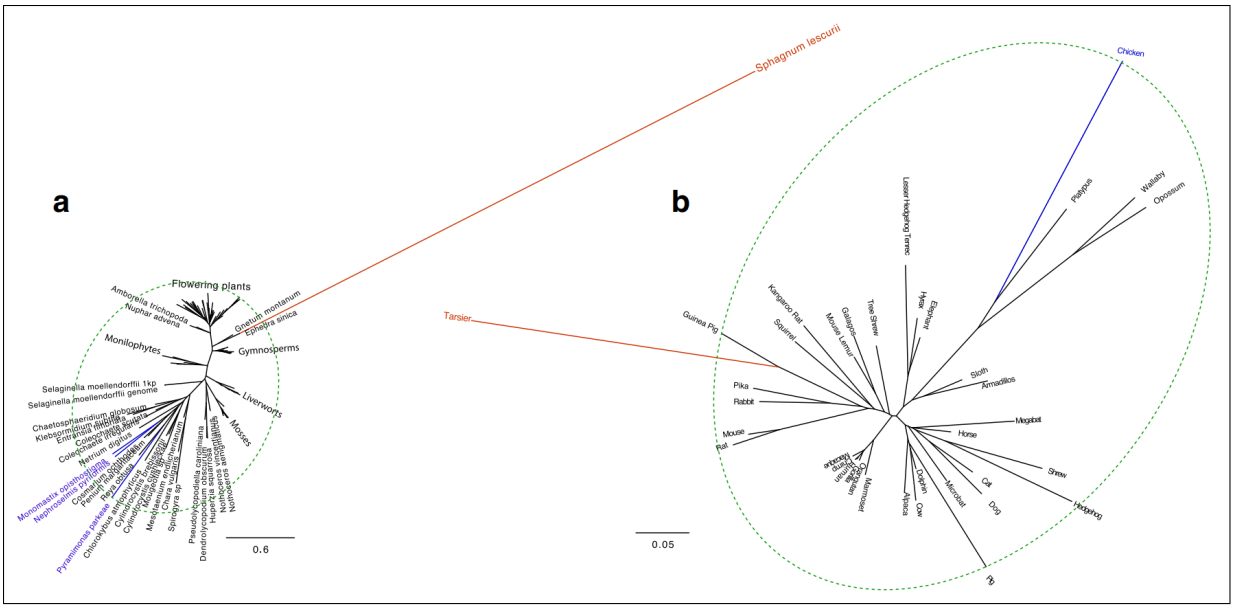

Figure 1: Typical trees with suspicious long branches. a. A gene tree in a plant dataset with a visible outlier leaf [25]; b. A gene tree in a mammalian dataset with a difficult-to-see outlier branch [20]. Outgroups are depicted in blue, and the long suspect branches are represented in red. The Green dashed line indicates the tree's diameter following the removal of its red branches. On the left, finding the red branch is simple; on the right, it is more complex.

By rerooting the tree at the centroid edge, the suggested approach would solve the issue in $O(k^2h+n)$ time, where h is the tree's height. In light of the k-shrink problem's solution, it is necessary to choose the species to eliminate to reduce the number of eliminated error-free sequences. As a result, three statistical tests are suggested to identify outlier species. In order to find outlier values in the spectrum of these proportional reductions, it is necessary to compute the diameter's proportional reduction as such moves from $i-1$ to i removals for $1 \leq i \leq k$. Outliers are defined as values that lie at the extreme tails of the distribution, and a level of false positive tolerance regulates the detection of outliers. Another difficulty is that outgroups can significantly affect the diameter even when error-free (figure 1b). For the second test, outliers in a single distribution are discovered by combining data from all gene trees. The final test takes things further by learning a new allotment for each species.

2 K-shrink problem solution in polynomial time

Two reasonable removals exist if t is singly paired, and one of them could reduce the diameter further. Given that this is readily verified, the problem is straightforward for $k = 1$. As a result, to overcome this problem, a search space must be considered. However, a brute force search for all plausible k-removing chains is impractical. The brute force technique would initially take into account the original diameter pair(s); after that, it would take into account the new diameter pair(s) after the first removal and recurse on each diameter pair to remove each of the two on-diameter leaves. All logical removing chains from 1 to k are produced by this recursive method, although its space expands exponentially.

Corollary 1: The proposed algorithm solves the k-shrink problem in $O(k^2h + n)$.

2.1 Application to all trees in general

Nodes in the search graph may contain more than two offspring if the tree t is not singly paired, which expands the search space. However, that may arbitrarily break the ties at any stage and yet ensure an ideal outcome. The algorithm logically also functions for trees that are not singly paired.

It is defined that a pair restricted k-removing space as a subset of $S_{\mathrm{k}}(t)$ such that each of its elements includes either a or b for each diameter pair $(a, b) \in P(t)$.

Theorem 1: There is at least one optimum k-removing set in any arbitrary pair-restricted k-removing space for any k.

The tree diameter stays constant unless all but one of the diameter groupings are eliminated. As a minimal shrunk tree of t, it is referred to as the limited tree of t with all but one of the diameter groups deleted. Suppose k is so tiny that there is no k-removing set that can reduce the tree diameter. In that case, any solution is optimal, and the result statistical selection of the filtering species of Theorem 1 trivially follows. Otherwise, any optimal solution of k-shrink can be induced from one of the minimum shrunk trees. Thus, to find an optimal tree t^*, it has to start from any pair-restricted removing space and concatenate the two removing chains: the chain that induces the minimum shrunk tree $t_{\mathrm{i}}{}^*$ from any arbitrary diameter pair, and the chain that starts from $t_{\mathrm{i}}{}^*$ to induce t^*. Any solution is optimum, and Theorem 1's conclusion follows if k is so little that no set that removes k from the equation may lower the tree's diameter. Alternatively, any k-shrink optimum solution may be inferred from one of the minimal shrunk trees. Therefore, to discover the best tree, t^*, it must start from any pair-restricted removing space and concatenate the two removing chains: the chain that starts from any arbitrary diameter pair to induce the minimal shrunk tree, $t_{\mathrm{i}}{}^*$, and the chain that begins from $t_{\mathrm{i}}{}^*$ to induce t^*.

This guarantees that the algorithm still correctly finds an optimal solution in $O(k^2h + n)$ when the search space size increases with $O(k^2)$.

3 Statistically chosen species serve as the filter (empirical statistics)

Let delta i equal to $\log(v_{\mathrm{i}})$, and let $\mathrm{v_i}$ be the ratio of the minimum diameter with i - 1 leaf removed to the minimum diameter with i leaves removed. The predicted i values for a tree without outlier branches are close to one (for instance, T1 in figure 2). It is anticipated that v_1 will be significantly greater than other i values for a tree with one outlier leaf on a very long branch (T2 in figure 2). It is anticipated that two species on a very long branch will have small values for v_1, huge values for v_2, and small values once again for $i > 2$. (T3 in figure 2). One with three species and the other with five species on two particularly long branches, it is anticipated that v_3 and v_8 will be significant, and other values will be minor (T4 in figure 2). Outliers were found using v values, but first, the idea of a signature must be explained.

The $\mathrm{v_i}$ values for the deleting sets that contain a species gauge how the species affects the tree's diameter. The highest $\triangle$i among all eliminating sets i that contain a species will be referred to as the signature of that species. Three separate methods were developed to determine what constitutes an extremely large. While the other two need a collection of gene trees, the first test may be performed on a single tree.

The "per-gene" test

Given this case's limited amount of data, a parametric technique was utilized to fit a log-normal distribution to the signatures. Values with a CDF over 1 - α were considered outliers, given a false positive tolerance rate α. The species connected to the outlier signatures are then eliminated.

The "all-gene" test

One distribution was made using the combined gene signature. A kernel density function was calculated for the empirical distribution of the total set of signature values [19]. Silverman's rule of thumb smoothing bandwidth and Gaussian kernels were used to estimate the density [19] (as implemented in the R package [23]).

The "per-species" test

The tree's diameter can come from outgroups and incorrect species (figure 1b). For each species' signature values, a non-parametric distribution is first constructed. When a species' signature for a gene is not known, zero is utilized as the gene's signature. Next, a threshold for the signature value corresponding to the selected α is computed for each species using the same non-parametric method as in the all-gene test. Each species is then excluded from the genes where its signature strictly exceeds the threshold for that species.

3.1 The settings of TreeShrink by default

The two parameters for TreeShrink are k and α. The default was set α to 0.05 (but users can choose other thresholds). Large values of k are not appropriate for the search for outlier species and may provide false results. A method was obtained that is fast enough for high n by choosing a value of k that increases sublinearly with n. Although the user must finally decide, $k = \min(\frac{n}{4}, 5\sqrt{n})$ was put as a default. This heuristic formula prevents setting k to numbers near to n while simultaneously ensuring that the running time does not increase less than quadratically with n.

Table 1. Summary of the biological datasets

Dataset	Species	Genes	Outgroups	Download
Plants [25]	104	852	Uronema sp., Monomastix opisthostigma, Nephroselmis pyriformis, Pyramimonas parkeae	DOI 10.1186/2047-217X-3-17
Mammals [20]	37	424	Chicken	DOI 10.13012/C5BG2KWG
Insects [15]	144	1478	Symphylella vulgaris, Glomeris pustulata, Lepeophtheirus salmonis, DAPHNIA PULEX, Cypridininae sp, Sarsinebalia urgorii, Celuca puligator, Litopenaeus vannamei, IXODES SCAPULARIS	http://esayyari.github.io/InsectsData
Cannon [5]	78	213	Monosiga brevicollis, Mnemiopsis leidyi, Pleurobrachia bachei, Euplokamis dunlapae, Salpingoeca rosetta	DOI 10.5061/dryad.493b7
Rouse [17]	26	393	Amphimedon queenslandica, Trichoplax adhaerens, Nematostella vectensis, Mnemiopsis leidyi	DOI 10.5061/dryad.79dq1
Frogs [6]	164	95	Protopterus annectens, Homo sapiens, Crocodylus siamensis, Gallus gallus, Ichthyophis bannanicus, Batrachuperus yenyuanensis, Andrias davidianus, Latimeria chalumnae	DOI 10.5061/dryad.12546.2

4 Evaluation procedures

Plants [25]: Early patterns of diversification within land plants and their related groups were established using this dataset of 104 plant/algae species and 852 genes. Transcriptoms provide the basis for the data. All gene trees derived from nucleotide data are begun in the analysis without the third codon position.
Insects [15]: There are 144 species and 1478 genes in this phylotranscriptomic collection of insects. This work estimated all 1478 gene trees using RAxML gene trees, whereas a different publication used ASTRAL to analyze species trees on the same dataset [18].
Metazoa-Cannon [5] and Rouse [17]: Xenacoelomorpha was identified as the sister taxon to Bilateria using the Cannon et al. dataset of 213 genes from 78 species collected throughout the animal tree of life. ASTRALII [13] was

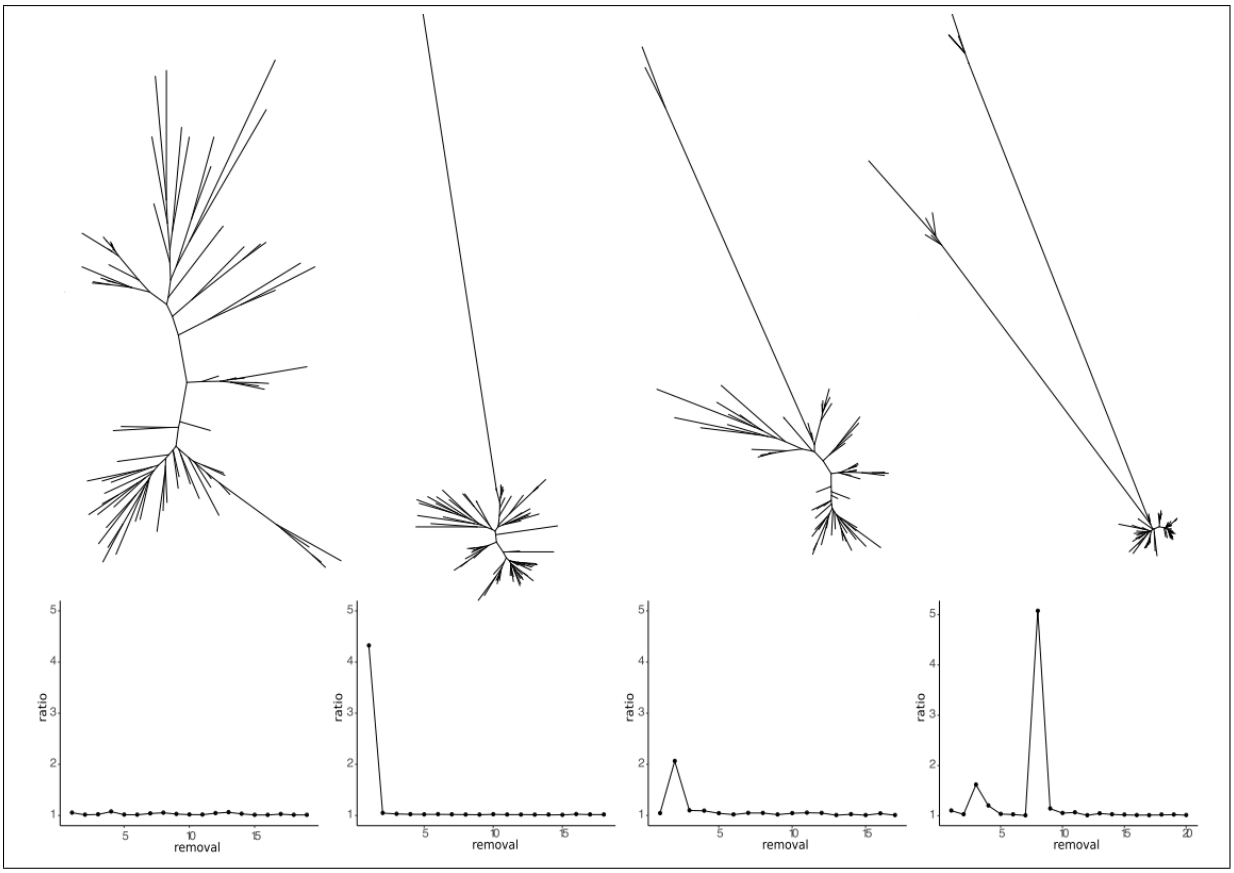

Figure 2: V_i's patterns in relation to i From a dataset of plants, four unfiltered gene trees [25] are shown (top). V_i is also demonstrated for $1 \leq i \leq k = min(\frac{n}{4}, 5\sqrt{n})$ for each tree (bottom).

employed on a set of gene trees that the authors disclosed and is utilized in this work, among other analyses.
Mammals [20]: There are 424 gene trees and 37 species of mammals in this collection. The study employed RAxML gene trees that were inferred and used to re-analyze this dataset [14] because the original gene trees had several problems.
Frogs [27]: This collection has 164 species and 95 genes. The authors [6] contributed the RAxML gene trees that were utilized in this work as inputs for ASTRAL to build the species tree.
HIV dataset [11]: 648 fragments of the HIV-1 pol sequences in this HIV dataset were utilized to recreate the neighborhood's HIV-1 transmission network from 1996 to 2011. The HIV-1 pol coding region dataset consists of 639 subtype B, seven non-subtype B, and two unassigned sequences. The authors provided the sequences, which have GenBank accession codes ranging from KJ722809 to KJ723456.

4.1 Tested Methods

The outgroup genes were included in 681 genes for the Plant dataset, and for the remaining genes, a linear-time version of the midpoint rooting was performed [12]. Each gene tree contained at least one of the outgroups in previous datasets. The study's approach is contrasted with RogueNaRok [1], which defines a rogue taxon as one that exhibits unstable locations in repeated bootstrap runs, even if the aims of RTR and our technique are slightly dissimilar.

4.2 Evaluation

It is difficult to assess the performance of the filtering techniques on actual data since it is unclear whether a deleted sequence is, in fact, inaccurate. Discordant patterns, meanwhile, can be useful. Erroneous sequences will worsen the perceived discordance even if actual gene trees may not match with the species tree. Therefore, good filtering should decrease the quantity of gene tree discordance between genes, and, arguably, more successful filtering techniques do so more so than less effective ones. So long as an optimization problem's goal is not to directly lower discordance, the effectiveness of a filtering technique may be assessed by how it affects gene tree discordance. Notably, none of the examined solutions attempts to lower the gene tree discordance directly or uses the species tree as an input. As a result, one metric for accuracy is the decline in discordance. The MS (Matching Splits) metric [2], implemented in the TreeCmp [3], is used to compare all pairings of gene trees and calculate gene tree discordance. Random removal is used as a control in order to make it easier to comprehend MS, which is not standardized.

For downstream analysis like species tree estimation, removing the same species from several genes might be much more challenging.

A knob on filtering techniques controls how much filtering is done. Various knob settings are investigated to prevent the effects of arbitrary decisions. α is set to 20 distinct values in the range [0.005, 0.1] for the three TreeShrink tests. By changing the weight factor for RogueNaRok to 21 values in the range [0, 1.0], it is possible to modify the penalizing factor of the dropset size. For rooted pruning, the number of standard deviations is changed between 0.25 and 5.00, in increments of 0.25, above the average representing long branches. The number of leaves is deleted in random pruning for each TreeShrink threshold on each gene tree, but the species is determined randomly.

The effectiveness of TreeShrink ($\alpha = 5\%$), rooted pruning (3 std), and RogueNaRok (default parameters) in spotting outliers are assessed on the HIV dataset.

4.2.1 First Experiment

The 648 sequences are used as the input for TreeShrink, which produces a RAxML tree. For rooted pruning, the RAxML tree is rooted at its halfway. One hundred bootstrap trees were made using RAxML to run RogueNaRok. The ability of TreeShrink rooted pruning and RogueNaRok to identify the seven non-subtype B and two unassigned sequences were examined.

4.2.2 Second Experiment

TreeShrink and rooted pruning were used to find the ten simulated outliers among the 639 subtype B sequences. Ten sequences were randomly chosen among the 639 subtype B sequences, and a tiny portion of their sites was altered to a random nucleotide randomly selected from the distribution of the base frequencies inferred from the original sequences. The three subtype C sequences were used to root the tree and were removed before feeding it to TreeShrink or rooted pruning. The tree was then rooted on the branch dividing the two subtypes. A total of 20 repetitions were

Table 2: Effects of the three TreeShrink tests on taxon occupancy.

Dataset	Method	Portion of data removed(%)	Portion of outgroups removed(%)
Plants	Per-gene	3.3	29.9
	All-gene	2.5	12.8
	Per-species	4.9	5.1
Mammals	Per-gene	0.6	11.8
	All-gene	1.2	17.0
	Per-species	3.6	4.7
Cannon	Per-gene	1.4	6.2
	All-gene	1.3	4.7
	Per-species	3.5	5.0
Rouse	Per-gene	1.3	1.9
	All-gene	1.2	1.1
	Per-species	4.0	4.5
Insects	Per-gene	1.2	6.6
	All-gene	0.8	2.9
	Per-species	4.3	5.0
Frogs	Per-gene	1.3	26.7
	All-gene	0.8	15.9
	Per-species	2.7	4.5

used to construct two sets of data, one with 5% of the sites modified and the other with 10%. FastTree provided estimates for the trees used in this experiment.

5 Results

5.1 Filtering's effects on taxon occupancy

Taxon occupancy is affected by the three TreeShrink tests, particularly for outgroups. Outgroups inevitably affect the tree's diameter, but under ideal circumstances, they shouldn't be lost more frequently than other leaves. The per-gene and all-gene tests tend to eliminate outgroups in all six datasets aggressively, but the per-species test nearly consistently eliminates all species, including outgroups (Table 1).

The Mammalian dataset's lone outgroup and worst case is chicken. The all-gene test removes chicken from 17% of the genes and the per-gene test from 12%. On the other hand, the per-species test is slightly more frequent than the average: it eliminates chicken in around 5% of the genes that have it and about 4% of the total data. There is some evidence, nevertheless, that the platypus is frequently misplaced in many of the gene trees in this dataset [21].

5.2 Filtering's effects on gene tree discordance

In order to reduce gene tree discordance with the least amount of filtering, the three TreeShrink tests are compared. When a method decreases the discordance more for a given amount of filtering, it is favored. All three tests of TreeShrink, except for the Frogs dataset, perform on average better than the random reference pruning. Overall, the per-species test consistently outperforms all other tests, followed by the per-gene test. There are significant differences between the per-species test and the all-gene tests for plants, mammals, and frog datasets, whereas there are less obvious differences for other datasets. For phylogenomic datasets with several genes, the per-species test of TreeShrink is recommended because it consistently performs the best here.

6 HIV Dataset

6.1 Identification of non-subtype B sequences

All seven non-subtype B sequences are successfully detected by TreeShrink using the input RAxML tree of the 648 HIV pol sequences. Importantly, TreeShrink leaves all subtype B sequences intact. Nevertheless, only one of the 41 rogue sequences identified by RogueNaRok is not a class B sequence. These variances result from the two techniques' dissimilar goals. Midpoint rooting allows for root pruning, which specifies three non-subtype B sequences as outliers while missing the other four and producing two false positives.

6.2 Identification of simulated outliers

Table 3. TreeShrink's effectiveness at detecting simulated outliers

Dataset	Method	True positives	False positives	Precision	Recall (Sensitivity)	Specificity
5% changed	TreeShrink	106	9	92.2%	53.0%	98.6%
	Rooted pruning	131	17	88.5%	65.5%	97.3%
10% changed	TreeShrink	198	0	100%	99.0%	100.0%
	Rooted pruning	200	0	100%	100.0%	100.0%

TreeShrink properly discovers 198 out of 200 outliers on the dataset with outliers at 10% altered in sequences, and rooted pruning correctly detects all 200 outliers; neither approach generates a false positive. TreeShrink properly discovers $\frac{106}{200}$ outliers with nine false positives on the dataset with outliers at 5% altered in the sequences, whereas rooted pruning detects $\frac{131}{200}$ outliers with 17 false positives. Table 2 and 3 shows that TreeShrink is a more cautious technique than rooted pruning since it has better accuracy and specificity but lower sensitivity.

7 Conclusion

In this study, TreeShrink is presented, a technique for removing species that have an unbalanced effect on the diameter of a phylogenetic tree. In phylogenomic datasets, TreeShrink efficiently reduces gene tree discordance and accurately identifies HIV subtypes. TreeShrink can be a new addition to an analytic pipeline as a supplement to cutting-edge rogue taxon removal techniques for screening sub-types, filtering contaminants, and detecting paralogs.[**knuthwebsite**]

References

[1] A. J. Aberer, D. Krompass, and A. Stamatakis. "Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice". In: *Syst Biol* 62.1 (2013), pp. 162–6. URL: `https://doi.org/10.1093/sysbio/sys078`.

[2] D. Bogdanowicz and K. Giaro. "Matching Split Distance for Unrooted Binary Phylogenetic Trees". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.1 (2012), pp. 150–60. URL: `https://doi.org/10.1109/TCBB.2011`.

[3] D. Bogdanowicz, K. Giaro, and Wróbel B. TreeCmp. "Comparison of trees in polynomial time. Evol Bioinforma. 2012;2012(8):475–87. 10.4137/EBO". In: *S* 9657 (). URL: `https://doi.org/`.

[4] M. J. Braun, J. E. Clements, and Gonda MA. "The visna virus genome: evidence for a hypervariable site in the env gene and sequence homology among lentivirus envelope proteins". In: *J Virol* 61.12 (1987), pp. 4046–54.

[5] J. T. Cannon et al. "Xenacoelomorpha is the sister group to Nephrozoa". In: *Nature* 530.7588 (2016), pp. 89–93.

[6] Y. Feng et al. "simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous–Paleogene boundary". In: *Dryad Digital Repository* 2546.2 (2017), p. 1. DOI: http://dx.doi.org/10.5061/dryad.12546.2. URL: https://doi.org/10.5061/dryad.

[7] P. A. Goloboff and Szumik CA. Identifying unstable taxa. "Efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok". In: *Mol Phylogenet Evol* 88 (2015), pp. 93–104. URL: https://doi.org/10.1016/j.ympev.2015.04.003.

[8] P. A. Hosner, E. L. Braun, and Kimball RT. "Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves: Galliformes: Odontophoridae)". In: *J Biogeogr* 42.10 (2015), pp. 1883–95. URL: https://doi.org/10.1111/jbi.12555.

[9] G. Jordan and N. Goldman. "The effects of alignment error and alignment filtering on the sitewise detection of positive selection". In: *Mol Biol Evol* 29.4 (2011), pp. 1125–39.

[10] D. Kr"uger and A. Gargas. "New measures of topological stability in phylogenetic trees - Taking taxon composition into account". In: *Bioinformation* 1.8 (2006), pp. 327–30.

[11] S. Little et al. "Using HIV networks to inform real time prevention interventions". In: *PLoS ONE* 9.6 (2014), p. 0098443. URL: https://doi.org/10.1371/journal.pone.

[12] U. Mai, E. Sayyari, and S. Mirarab. "Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction". In: *PLoS ONE* 12 (2017), p. 8. URL: https://doi.org/10.1371/journal.pone.0182238.

[13] S. Mirarab and Warnow T. Astral-ii. "coalescent-based species tree estimation with many hundreds of taxa and thousands of genes". In: *Bioinformatics* 31.12 (2015), pp. 44–52. URL: https://doi.org/10.1093/bioinformatics/btv234.

[14] S. Mirarab et al. "Statistical binning enables an accurate coalescent-based estimation of the avian tree". In: *Science* 346 (2014), p. 6215. URL: https://doi.org/10.1126/science.1250463.

[15] B. Misof et al. "Phylogenomics resolves the timing and pattern of insect evolution". In: *Science* 346.6210 (2014), pp. 763–67.

[16] N. D. Pattengale, K. M. Swenson, and Moret BME. "Uncovering hidden phylogenetic consensus". In: *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6053 LNBI. , Heidelberg: ; p.* Berlin: Springer, 2010, pp. 128–39.

[17] G. W. Rouse et al. "New deep-sea species of Xenoturbella and the position of Xenacoelomorpha". In: *Nature* 530.7588 (2016), pp. 94–97.

[18] E. Sayyari, J. B. Whitfield, and S. Mirarab. *Fragmentary gene sequences negatively impact gene tree and species tree reconstruction.* Mol Biol EvolIn press, 2017.

[19] B. Silverman. "Density estimation for statistics and data analysis". In: *Monographs on Statistics and Applied Probability.* London: Chapman Hall; 1986.

[20] S. Song et al. "Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model". In: *Proc Natl Acad Sci* 109.37 (2012), pp. 14942–7. URL: https://doi.org/10.1073/pnas.

[21] M. S. Springer and J. Gatesy. "The gene tree delusion". In: *Mol Phylogenet Evol* 94 (2016), pp. 1–33. URL: https://doi.org/10.1016/j.ympev.2015.07.018.

[22] J. W. Streicher, J. A. Schulte, and Wiens JJ. "How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards". In: *Syst Biol* 65.1 (2016), pp. 128–45. URL: https://doi.org/10.

[23] R. Core Team. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing; 2016.

[24] K. M. Westover et al. "biological companion to simulation analysis". In: *Mol Phylogenet Evol* 69.1 (2013), pp. 1–3. URL: https://doi.org/10.1016/j.ympev.2013.05.010.

[25] N. J. Wickfett et al. "Phylotranscriptomic analysis of the origin and early diversification of land plants". In: *Proc Natl Acad Sci* 111.45 (2014), pp. 4859–4868. URL: https://doi.org/10.1073/pnas.1323926111.

[26] Shen X-x, C. T. Hittinger, and A. Rokas. "Studies Can Be Driven By a Handful of Genes". In: *Nature* 1 (2017), pp. 1–10. URL: https://doi.org/10.1038/s41559-017-0126.

[27] Feng Y-j et al. In: *simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary* 114.29 (2017), pp. 5864–70. URL: https://doi.org/10.