A Project Report on USA

Sarthak Mishra*[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign

December 7, 2021

Abstract

Inferring a species tree is very challenging. And several studies has shown that site-based methods such as SVDquartets perform better than summary-based methods, such as ASTRAL when the Gene tree estimation error and Incomplete lineage sorting is very high. We present a method Utilizing SVDquartets on ASTRAL, USA, which uses SVDquartets to improvise the RF error of ASTRAL. We have tested our method on the simulated dataset used by Molloy and Warnow in their study.

1 Introduction

Inferring the species tree is a very captivating work. Many biological processes, such as speciation hybridization, cause different regions of genomes to evolve differently [10]. One such process is Multi-species coalescent [10]. Under MSC, the gene evolves within the species tree and can be different from the species tree in "Incomplete lineage sorting." The MSC treats each species as s population of individuals. And each of these individuals has a set of alleles for each gene. Over time different alleles assort into different populations so that speciation events can lead to different species having different sets of alleles among their individuals[10]. When this happens, a gene tree defined using a single allele from each selected individual can differ from the species tree and each other. The forwards described are called "lineage sorting," If this process results in tree discordance, it is called incomplete lineage sorting [10]. ASTRAL is a species tree estimation method that takes a set of gene trees as the input-output of the species tree. ASTRAL solves the maximum quartet support subtree problem to get the species tree[5, 12]. SVDquartets is a site-based species tree estimation method that takes the unlinked loci and estimates the species tree from the distribution of the site patterns[10]. SVDquartets uses Single value

*mishra20@illinois.edu

Decomposition to infer the quartets and uses Quartet Max cut or quartet FM to combine the quartets[1]. The summary methods, such as ASTRAL, depend upon the gene tree estimation to prove statistical consistency. If the input is the true gene tree, then the gene tree estimation error is zero. But as the gene tree estimation error increases, the summary methods such as ASTRAL will perform poorly[10]. The study conducted by Molloy and Warnow [6] shows that the site-based method such as SVDquartets performs better than summary methods ASTRAL when the gene tree estimation error is super high, and the Incomplete linear sorting is very high[6]. For our course project, we have used the study by Molloy and Warnow [6] to come up with a plan to improvise ASTRAL using SVDquartets for a high Gene Tree Estimation Error dataset. This study shows that SVDquartets perform much better than ASTRAL in certain alignment data sets that have very high Gene Tree Estimation Error. We are picking these replicates and running ASTRAL in them. After this, we collapse low support branches and run SVDquartets to resolve the polytomy.

2 Materials and Methods

2.1 Dataset

The data used in our experiment was obtained from Molloy and Warnow[6]. The data used was ASTRAL-II dataset. Then this data was simulated using SimPhy to evolve genes with Multispecies coalescent (MSC) model. The original data had two speciation rates 10^6 and 10^7, three different tree heights(10M, 2M, 500K), 1000 genes, and 200 species. But they reduced the number of species to 26 to study the computationally intensive MP-EST[6]. Our study uses model conditions with a 10^6 speciation rate, 500k tree height, 26 species, and five different gene numbers. These five numbers are 50, 100,250,500, 1000. We assembled these genes number going sequentially in the file. So, for instance, the data with 50 genes was created by going over each file from 0001. fas to 0050. Similarly, data with 100 genes was created by going over the files from 0001.fas to 0100.fas, 250 was created using gene files from 001.fas to 250.fas. 500 with 001.fas to 500.fas and 1000 with 0001.fas to 1000.fas .Molloy and Warnow [6] point out that the tree with shorter heights have higher Incomplete lineage sorting, and they have also provided the replicates with very high Gene Tree Estimation Error. These replicates are in Table S6 of their supplement. These replicates are estimated from the short sequence (100 sites), and the mean gene tree estimation error estimated in the Robinson foulds (RF) error [7] rate between the true gene trees and the estimated gene trees is 80% for these datasets. Molloy and Warnow [6] study have also pointed out that SVDquartets are more accurate than ASTRAL for datasets with a mean Gene tree estimation error of 80% or more. Hence we are using 5 of these replicates from species tree height of 500k, speciation rate of 10^6 which are 05,11,13,16,20 [6].

2.2 Methods

We are comparing our method against ASTRAL- a summary-based method [5, 12] and SVDquartets- a site-based method [1]. ASTRAL gives branch support values for all the computed branches. A branch support value measures the quadripatition around a branch[5, 12]. We will be collapsing branches whose support is less than certain threshold values. For our study, we have 0.35,0.4 and 0.8 as the threshold values to collapse the branch with support values less than the selected threshold. And we create polytomies and resolve them using SVDquartets. This is the main flow of our method USA.

2.3 Benchmark Criteria

The testing statistic that we will use in our experiments is the RF distance [7] in order to gauge species tree error similar to the experimental test statistics run in Molloy and Warnow [5]. We also tested the run time for all methods involved. We used Dendropy [9] to compare the trees and compute the statistic.

3 USA: Utilizing SVDquartets on ASTRAL

We present our method Utilizing SVDquartets on ASTRAL, USA, which runs ASTRAL, collapses the low support branches, creates polytomies, and resolves these polytomies using SVDquartets. The input for our method is Alignments which is used to compute gene trees by RAxML [8], the gene trees are then used to run run ASTRAL, and the output is the improvised species tree using SVDquartets. Below we explained the details of the pipeline of our method. Initially we decided not to run RAxML and use the gene trees provided by Molloy and Warnow [6] to run ASTRAL.In doing so we ran into a situation where some taxa that were present in species tree given by ASTRAL were not present in the alignment file. Computing the gene tree using RAxML solved this issue.

We engineered the pipeline in the following way. In the first step, we ran the RAxML[8] to get the estimated gene trees for each locus of the gene. The RAxML was run on only the first 100 sites. On these different gene trees given by RAxML, we then run the ASTRAL. Then the NEWICK utilities[3] was used to collapse the low branch support edges. There were three different thresholds used for collapsing the branch support. Then we traverse the collapsed tree using Dendropy[9] and collect all the polytomies. Then for each polytomy, we select one taxon from each subtree then extract the corresponding alignments from the replicates, and these were further converted to nex files. After this process, all the nex files were combined to form a larger nex file. This file was used to run PAUP*+SVDquarters [11]. This was done for each polytomy.

We then regrafted the tree given by the SVDquartets to the collapsed tree. In doing so, firstly, we replaced the taxon in the tree given by SVDquartets with the corresponding subtree from the collapsed tree. We do it for all the taxa. Once all the taxa are replaced, we regraft this tree to the collapsed tree. Again

we used Dendropy[9] to compare the re-graftred tree with the true species tree given for each replicate. The Robinson fouled [7] distance was used to compare the true tree with the USA's tree.

4 Results

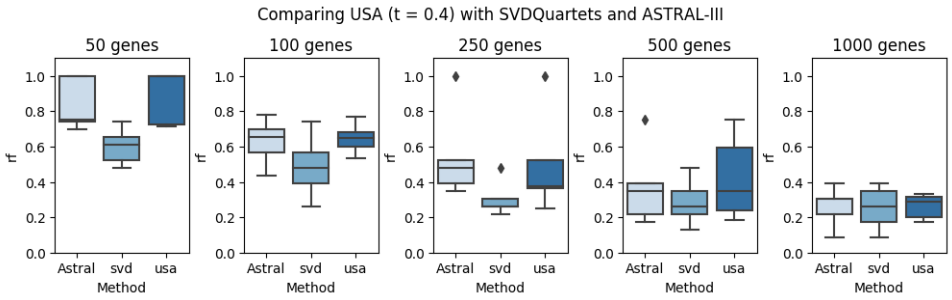

Figure 1: Comparing the RF error of our method USA with ASTRAL and SVDquartets for the threshold value of 0.4 where one taxon from each subtree in a polytomy was selected

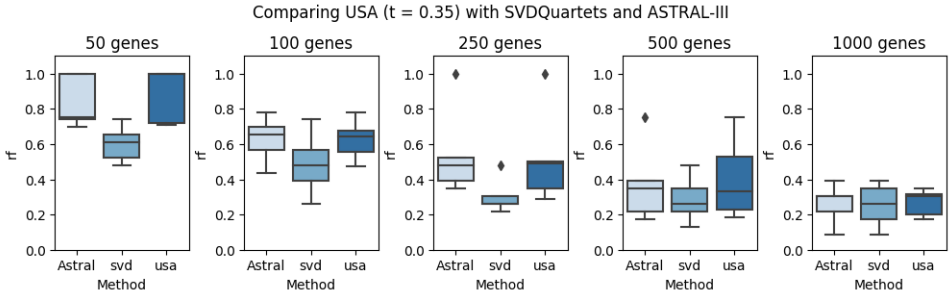

Figure 2: Comparing the RF error of our method USA with ASTRAL and SVDquartets for the threshold value of 0.35where one taxon from each subtree in a polytomy was selected

In figure 1 our method Utilizing SVDquartets on ASTRAL(USA) was compared with two other methods, ASTRAL, a summary-based method, and SVDquartets, a site-based method. When the threshold was 0.4, one taxon was selected from each taxon on subtree on the created polytomies. SVDquartets is the most accurate in all the cases except when using all 1000 genes. USA is equally accurate or more accurate than ASTRAL when there are missing genes. In

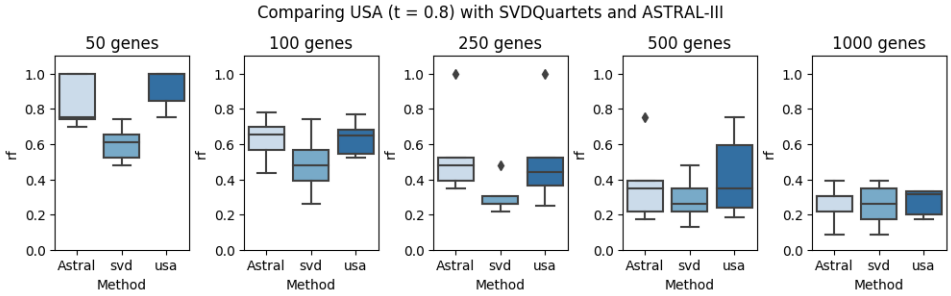

Figure 3: Comparing the RF error of our method USA with ASTRAL and SVDquartets for the threshold value of 0.8 where one taxon from each subtree in a polytomy was selected

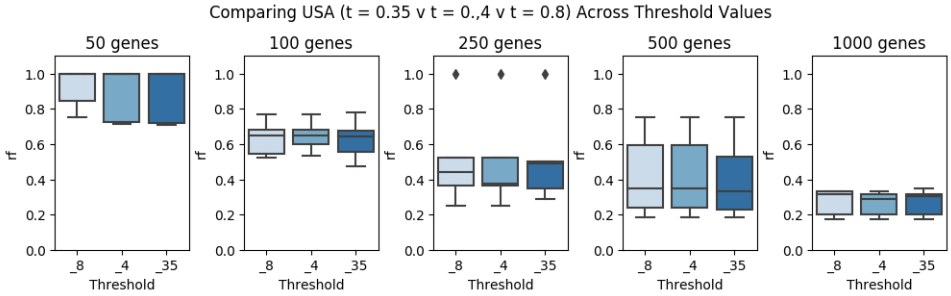

Figure 4: Comparing the RF error of our method for three different threshold value[0.35,0.4,0.8].

particular, when the number of genes is 250, our method USA is better than ASTRAL, but it is not as good as the USA.
In figure 2 our method Utilizing SVDquartets on ASTRAL(USA) was compared with two other methods, ASTRAL, a summary-based method, and SVDquartets, a site-based method. When the threshold was 0.35, one taxon was selected from each taxon on subtree on the created polytomies. SVDquartets is the most accurate in all the cases except when we are using all 1000 genes. USA is equally as accurate or more accurate than ASTRAL when there are missing genes. In particular, for the case when the number of genes is 250, our method USA is clearly better than ASTRAL, but it is not as good as the USA.
In figure 3 our method Utilizing SVDquartets on ASTRAL(USA) was compared with two other methods, ASTRAL, a summary-based method, and SVDquartets, a site-based method in terms of rf error. When the threshold was 0.8, one taxon was selected from each taxon on subtree on the created polytomies. SVDquartets is the most accurate in all the cases except when we are using

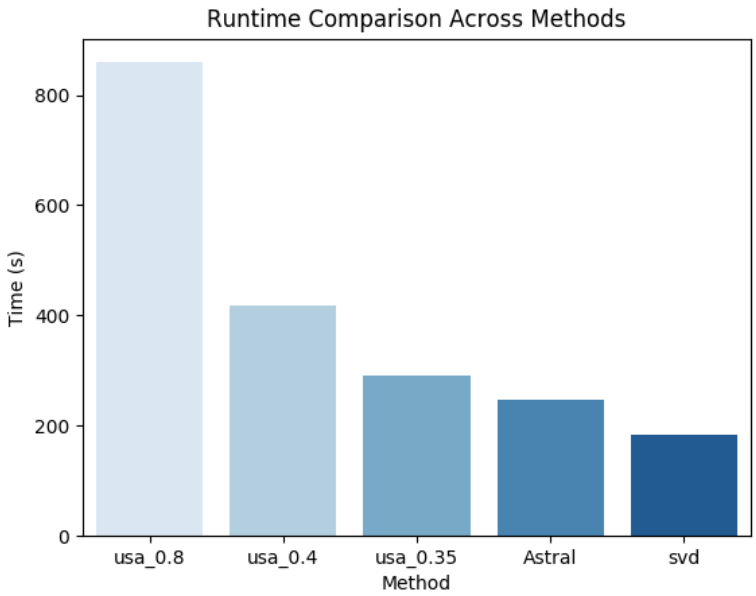

Figure 5: Comparison of running time in seconds for all the methods involved.

all 1000 genes. USA is equally as accurate as ASTRAL in some cases, such as when the number of genes is 250 .But it is not better than astral in most other conditions i.e, when the number of gene is 50,100,500 and 1000.
In Figure 4 shows the rf error across three threshold values. Our experiment shows that 0.35 is the best threshold value as rf error when the threshold is 0.35 is less than when the threshold is 0.4 or 0.8. The threshold value of 0.4 gives the best results when the number of genes is 1000 but lags behind 0.35 when there are a missing number of genes. Similarly, the threshold value of 0.8 is the worst in all numbers of genes.
In Figure 5 shows the running time in seconds for all the methods used. The running time for the USA is affected by the time taken by ASTRAL as we are running ASTRAL and collapsing its low support edges and running the SVDquartets. As the value of threshold increases, the running time of USA increases, the USA with 0.8 threshold value is the longest-running, whereas SVDquartets is the shortest running method.

5 Observation And Discussion

One of the observations we are seeing is that the threshold value of 0.35 is the best in terms of accuracy that takes a hit negatively when we increase the

branch collapse threshold value. This was not an expected result as we initially thought having a higher threshold means we are collapsing more edges, thereby creating more polytomy and resolving these with SVDquartes, which is shown to be more accurate than ASTRAL must decrease Robinson Fould (r.f) error. But these results show otherwise. One reason for this could be a scenario when the branch support value of an ASTRAL tree is lower in branches closer to the root, but it has a higher branch support value in the branches closer to leaves. Also, there could be a situation when the threshold value of 0.8 could be too large, and it could just collapse most of the branches from the ASTRAL tree. This means we are trying to run everything using SVDquartets but using only one taxon from a subtree which may have negatively affected the accuracy. In these cases, 0.35 could be ideally suited to decrease just a good number of branches to improvise the accuracy of ASTRAL on a high gene tree estimation error dataset.

Another observation is SVDquartets is the most accurate method in all the missing gene conditions. Although ASTRAL is better than SVDquartets and USA when we have all 1000 genes included, SVDquartets gets significantly better than ASTRAL and USA when the number of missing genes increases. Our technique of keeping flat-out threshold numbers for collapsing the branches may not be the best way to improvise the ASTRAL using SVDquartets. For some species of trees created by ASTRAL, the mean of overall distribution of branch support could be much lower, and for trees, it could be very high. Hence collapsing branches less than the fixed value may not be the best way to deal with the edge. For this, we think using something like a normal distribution to get the value of the threshold for each species tree could make this study much better. For example, we can calculate the mean and standard deviation of the branch support from each ASTRAL species tree and compute the threshold such that when the z value is 0.1. This will provide us with the threshold for each ASTRAL species tree rather than a flat-out threshold.

Another avenue to explore would be using some biological data such as Avian data by Jarvis et al.[2]. Since there were no indels in the Molloy and Warnow datasets, we did not use the multiple sequence alignments such as MAFFT [4]. But in the case of indel datasets, we can use MAFFT and use that alignment to run SVDquartets. This might improve the accuracy of the USA.

6 Conclusion

Our experiment concludes that there is a potential to improve ASTRAL using SVDquartets for a high gene tree estimation error dataset. Although we see some positive trends, our experiment is unable to determine if ASTRAL can be improvised by SVDquartets using our method USA. We have suggested some techniques, such as using the z score to get the threshold value for each ASTRAL species tree and collapse them according to that value. Also, our suggestion of expanding this study with other multi-locus datasets could help determine the validity of our method USA.

References

[1] Chifman, J., and Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics 30*, 23 (Dec 2014), 3317–3324.

[2] Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science 346*, 6215 (2014), 1320–1331.

[3] Junier, T., and Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics 26*, 13 (05 2010), 1669–1670.

[4] Katoh, K., Misawa, K., Kuma, K., and Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research 30*, 14 (07 2002), 3059–3066.

[5] Mirarab, S., and Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics 31*, 12 (06 2015), i44–i52.

[6] Molloy, E. K., and Warnow, T. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology 67*, 2 (09 2017), 285–303.

[7] Robinson, D., and Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences 53*, 1 (1981), 131–147.

[8] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*, 21 (08 2006), 2688–2690.

[9] Sukumaran, J., and Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics 26*, 12 (Jun 2010), 1569–1571.

[10] Warnow, T. *Phylogenomics: Constructing Species Phylogenies from Multi-Locus Data*. Cambridge University Press, 2017, p. 234–273.

[11] Wiens, J. J., Kuczynski, C. A., Smith, S. A., Mulcahy, D. G., Sites, Jack W, J., Townsend, T. M., and Reeder, T. W. Branch Lengths, Support, and Congruence: Testing the Phylogenomic Approach with 20 Nuclear Loci in Snakes. *Systematic Biology 57*, 3 (06 2008), 420–431.

[12] Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics 19*, 6 (May 2018), 153.