Group 30

Group Members: Bang Ly, Ethan Tenley, and Joshua O'Neill

Professor Dan Knights

CSCI 5481 | Computation Techniques for Genomics

11 December 2024

Final Project

Dynamic Grouping Algorithm

Abstract

Biodiversity assessment is necessary for understanding ecosystem health and species interactions, and phylogenetic tools are often utilized to assess that diversity. Traditional unweighted phylogenetic algorithms calculate biodiversity based on tree topology alone, while weighted methods take into account species abundance, providing an additional perspective. However, weighted phylogenetic approaches have been shown to favor samples with high abundances of a few species, misrepresenting the diversity of evenly distributed samples. Here, we show an analysis of and improvement in weighted phylogenetic diversity calculations by introducing our dynamic grouping algorithm. This algorithm addresses the aforementioned problems by weighting edges based on the collective abundance of descendant groups, ensuring a more balanced and unbiased species distribution. Compared to the traditional weighted algorithm, our approach more effectively distinguishes diverse samples, addressing biases and aligning more with ecological expectations. Our findings highlight the limitations of conventional methods, demonstrating that weighting by individual abundances can obscure the true diversity. By utilizing group-level abundance distributions, our dynamic grouping algorithm enhances biodiversity assessment. This work has broad applications for ecological and evolutionary studies, as our algorithm provides researchers with an additional tool to analyze larger and more complex real-world biodiversity assessment datasets.

Summary of Previous Findings

The first findings come from the article "Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny"[1]. The paper talks about the limitations of using UniFrac and weighted UniFrac (W-UniFrac) where W-UniFrac weighted length of the branch by the difference of relative abundances of the two communities in that tree but it assumes that the two communities are the same. If the two communities are different, there would be a variance in weight for different branches. The article proposed a different weighting approach in the W-UniFrac by considering the weight variation called Variance Adjusted Weighted UniFrac (VAW-UniFrac). The results show that the VAW-UniFrac is more powerful than W-UniFrac. There is currently no limitation found in the paper aside from

more data containing an abundance of information is needed to help obtain new insights about community differences.

The second findings come from the article “Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth”[2]. The paper talks about the utility of abundance-weighted phylogenetic diversity (PD) which uses the sum of the weighted branch length of the tree. Then, it introduces a new generalized measure called Balance-Weighted Phylogenetic Diversity ($BWPD_0$). It used a parameter θ as a way to adjust and balance between classical PD and abundance-weighted PD. $BWPD_0$ is classical PD, whereas $BWPD_1$ is abundance-weighted PD. The result shows that Operational Taxonomic Unit OTU-based measures are less effective in distinguishing community types. $BWPD_{0.25}$ and $BWPD_{0.5}$ were the only measures that were in the top four for all datasets (Bacterial vaginosis, Oral periodontitis, and Skin microbiome through time datasets) due to them being less sensitive to different sampling intensities. The limitation would be finding the correct parameter θ.

As for the novelty of the project, VAW-UniFrac can only be used for beta diversity and $BWPD_0$ for alpha diversity. Both algorithms used the weighted length of the tree branches as a way to measure the abundance. The Dynamic Grouping algorithm can be used for both alpha and beta diversity. The algorithm penalizes branches through weights solely based on the unevenness of abundance and does not penalize when abundance is evenly distributed.

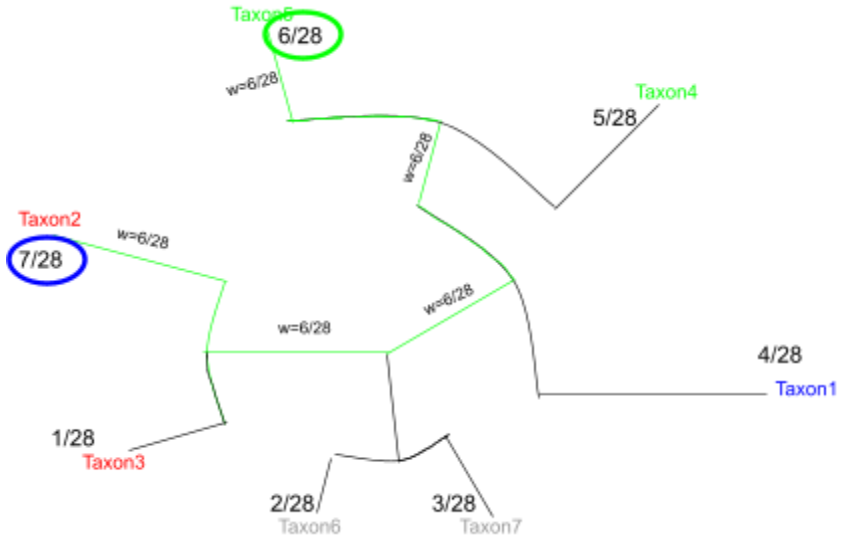
Results

The first approach used to find the biodiversity of a sample using weighted phylogeny was quite simple. Similar to an unweighted phylogenetic calculation, we start with an empty tree and then add edges and nodes to it until the sample is spanned by the tree. The first step to adding to this tree is finding the species with the maximum abundance and adding it to the tree. At this point, no edges are part of the tree; just the node representing the species found with maximum abundance (See figure 1-a). Then, the node with the second most abundance was found and added to the tree, and a line is drawn between this node and the rest of the tree. In this case, the rest of the tree is just the node/species with the most abundance. The path from this new node to the rest of the tree is added to a calculation of biodiversity, as occurs with an unweighted phylogeny algorithm, but unlike an unweighted algorithm, this sum of branch lengths is weighted by the abundance of the node being considered. In this case, the node being considered is the second most abundant node (See figure 1-b). Then, the third most abundant node is selected and added to the tree in a similar fashion as the second most abundant node (See figure 1-c). This process continues until there are no nodes left to add the tree, in which case the algorithm has finished (See figure 1-d).

(a)

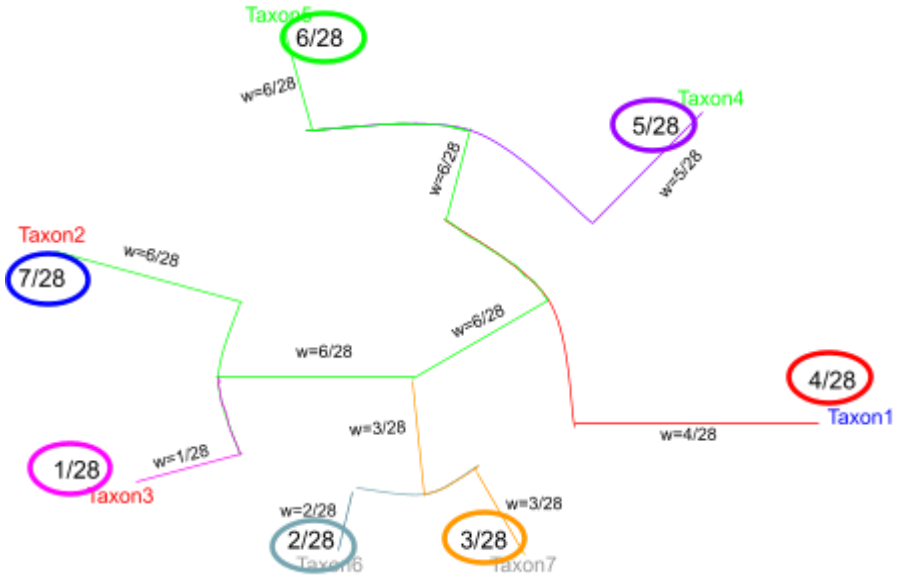

Taxon5
6/28
Taxon4
5/28
Taxon2
7/28
4/28
Taxon1
1/28
Taxon3
2/28
Taxon6
3/28
Taxon7

(b)


Taxon5
6/28
w=6/28
Taxon4
5/28
w=6/28
Taxon2
w=6/28
7/28
w=6/28
w=6/28
4/28
Taxon1
1/28
Taxon3
2/28
Taxon6
3/28
Taxon7

(c)


Taxon5
6/28
w=6/28
Taxon4
5/28
w=6/28
w=5/28
Taxon2
w=6/28
7/28
w=6/28
w=6/28
4/28
Taxon1
1/28
Taxon3
2/28
Taxon6
3/28
Taxon7



(d)

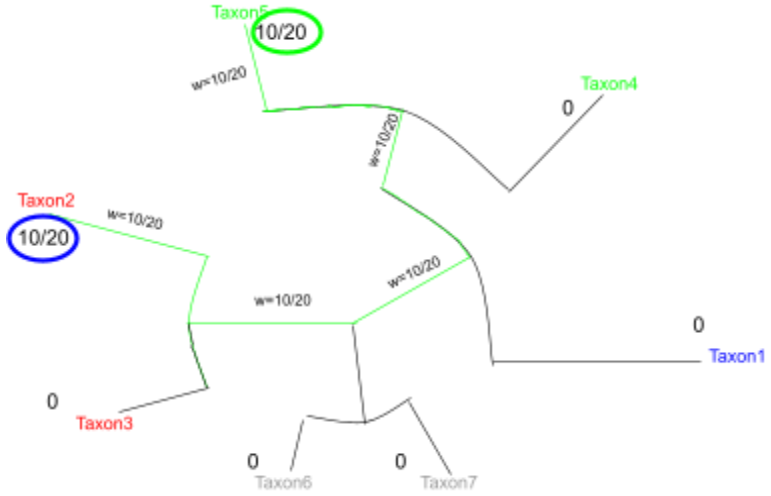
Figure 1: Shows the initial weighted phylogenetic diversity algorithm performed on a tree of 7 taxons. a-d show steps of the algorithm being performed.

After programming this algorithm and testing it on 12 samples from a 7-node tree, we discovered many flaws in this algorithm (table 1), namely that it rewards trees with few species with high abundances more than trees, such as with sample 2, 3, 4, 8, 9, and 11, with many species and low abundances, such as with sample 1 and 10, despite the latter being more diverse. By examining sample 11 and 12, we discovered the cause (Figure 2). The issue that occurs is that the branches are weighted by species abundance, meaning that a species with a lower abundance count will contribute less to the tree. This means that even though more of the tree is covered with more species, it is likely to be counted as less diverse due to less weight being applied to the tree. To see an illustration of this issue, please see figure 2.

Taxon1	Taxon2	Taxon3	Taxon4	Taxon5	Taxon6	Taxon7	Score
1/7	1/7	1/7	1/7	1/7	1/7	1/7	0.1884
3/10	1/10	0	3/10	3/10	0	0	0.2443
3/10	0	0	3/10	3/10	1/10	0	0.2303
3/10	1/40	1/40	3/10	3/10	1/40	1/40	0.2211
14/20	1/20	1/20	1/20	1/20	1/20	1/20	0.0660
1	0	0	0	0	0	0	0
0	1/4	1/4	0	0	1/4	1/4	0.1250
1/3	1/3	0	1/3	0	0	0	0.3354
1/3	1/3	0	0	0	1/3	0	0.3121
1/6	1/6	1/6	1/12	1/12	1/6	1/6	0.1987
0	1/2	0	0	1/2	0	0	0.2771
0	1/2	0	1/4	1/4	0	0	0.1609

Table 1. Each row of this table represents a test sample. Columns 1-7 list the abundance of each Taxon in each test sample, and column 8 lists the diversity score given to it by the initial algorithm explored by this paper.

(a)



(b)

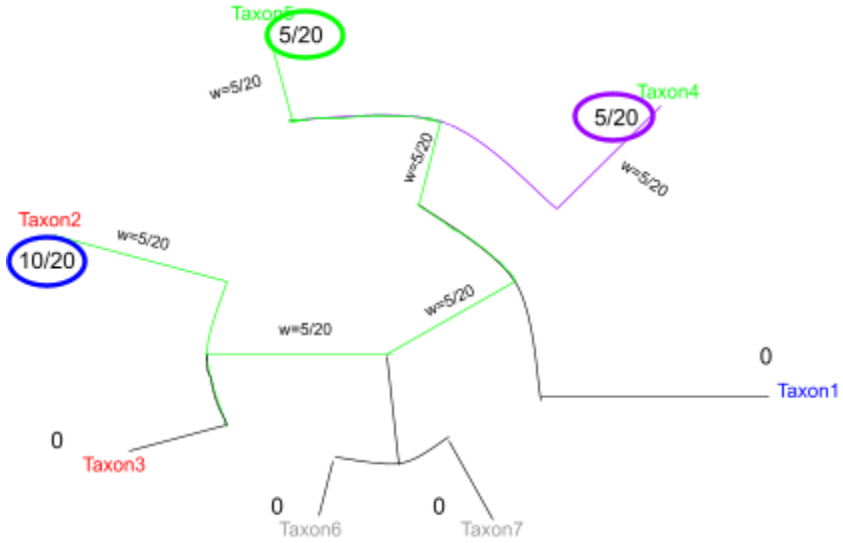


Figure 2. Compares two different samples using the initial algorithm from this paper to illustrate shortcomings. (a) High weight given to tree connection despite lacking species variety due to high abundance of singular species. (b) Significantly lower weight given to tree where abundances of one section of the tree are split up relative to (a) causing more diversity.

This discovery led to many changes. While it is important to attribute less weight to edges added from areas of the tree with low abundance, it's important to still promote regions with only nodes with low abundances, but an even distribution of low abundances that

collectively add up to have high abundance. This goal naturally leads to the idea of examining nodes based on groups. Each edge should be weighted not based on a node's abundance descending from it, but all the nodes descending from it (Figure 3). This way, edges will still be adequately weighted. To accomplish this, an entirely different approach than the aforementioned algorithm is required, and after experimenting with numerous approaches, an algorithm was created that mirrored the unweighted algorithm, except it penalizes branches through weights solely based on the unevenness of abundances. The idea is that, in a fully balanced tree, each child of the root will have descendants who comprise exactly ⅓ of the sample in total. Then below that, each child of these children will possess ½ of their parents abundance, so ⅙ of the total tree's abundance. This balance is ideal for all nodes on the tree for a fully balanced tree (Figure 3). This tree is one in which no penalty is applied, and so the tree has the same diversity as the unweighted calculation. The way this ideal is implemented is by adding nodes to the tree as was done in the initial approach, except once a node is encountered that has nodes already added to the tree as its descendants, the weights are determined by comparing both of this node's childrens' descendants' abundances to one another. Conceptually, this is a comparison between the new node's group's abundance, represented by the child with the new node and without the nodes already added to the tree as descendants, and the currently existing tree's abundance, represented by the child without the new node and with the nodes already added to the tree as descendants. The proportion between the two groups' abundances is then used as a penalty for all edges being added to the tree through the new node. This penalty is applied on top of any penalty incurred by the group it is connected to. A skewed part of a skewed tree is extra skewed, after all. Figure 4 shows a walkthrough of this algorithm. This updated algorithm has led to substantially more success than the initial version. It has proven successful in distinguishing itself both from the initial algorithm by favoring more even distributions of abundances and unweighted methods by treating parts of the tree with collectively low abundances as relatively insignificant (Table 2).

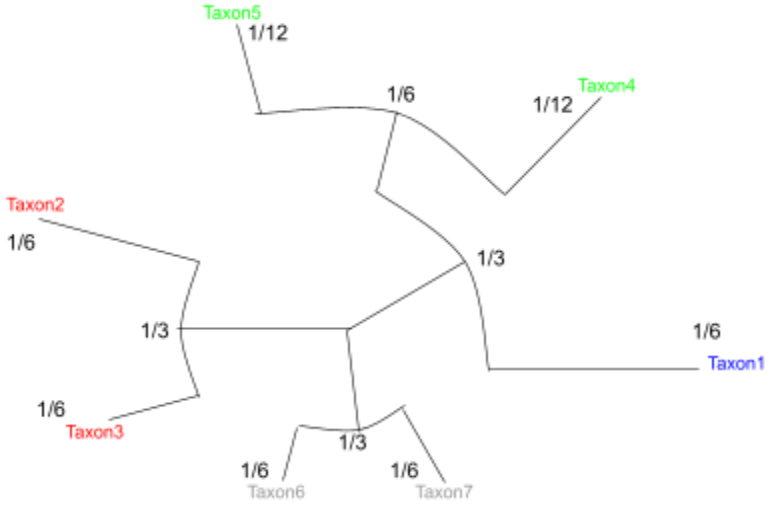
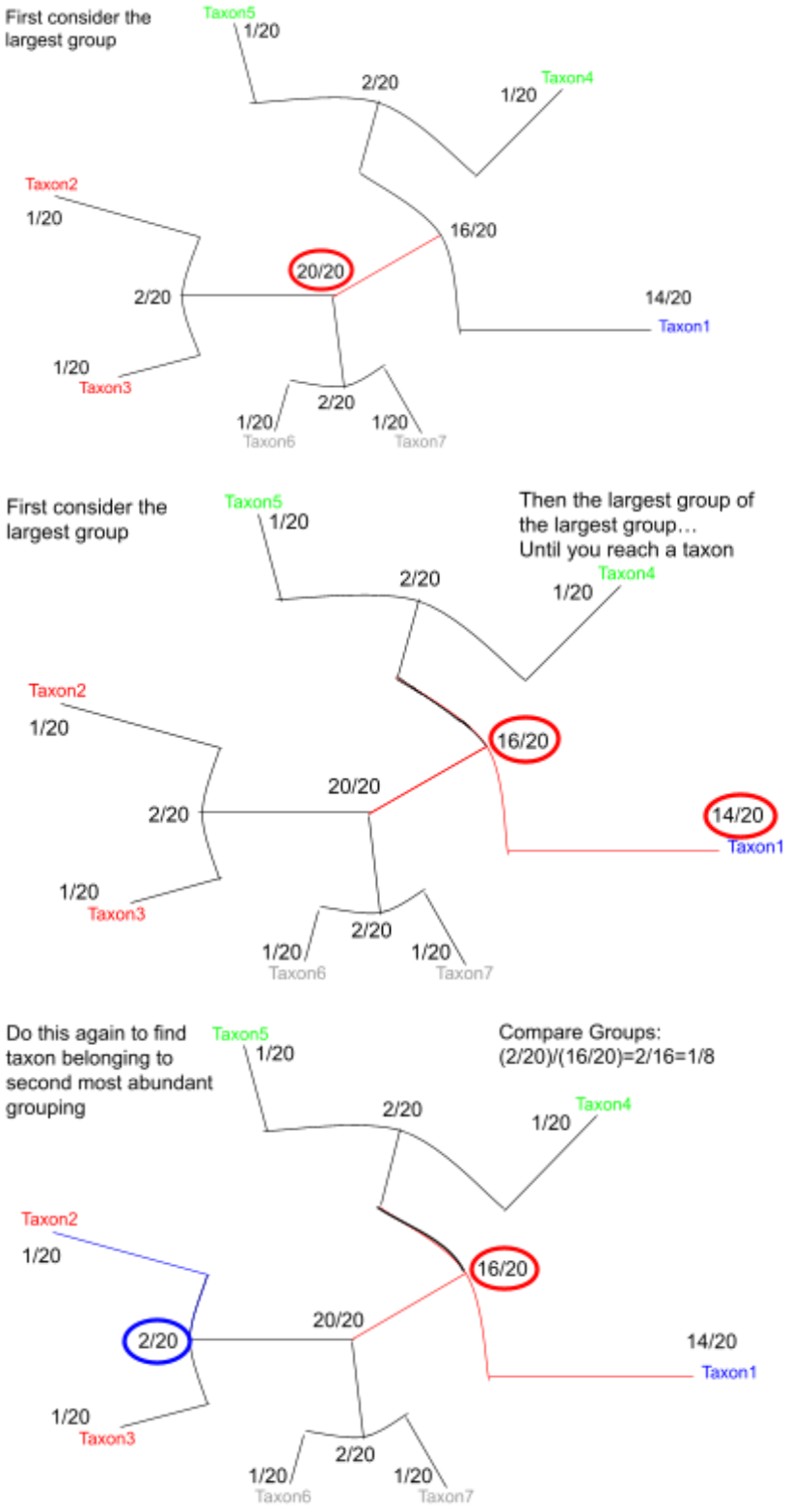


Figure 3. Shows a fully balanced tree based on the abundance of groups.



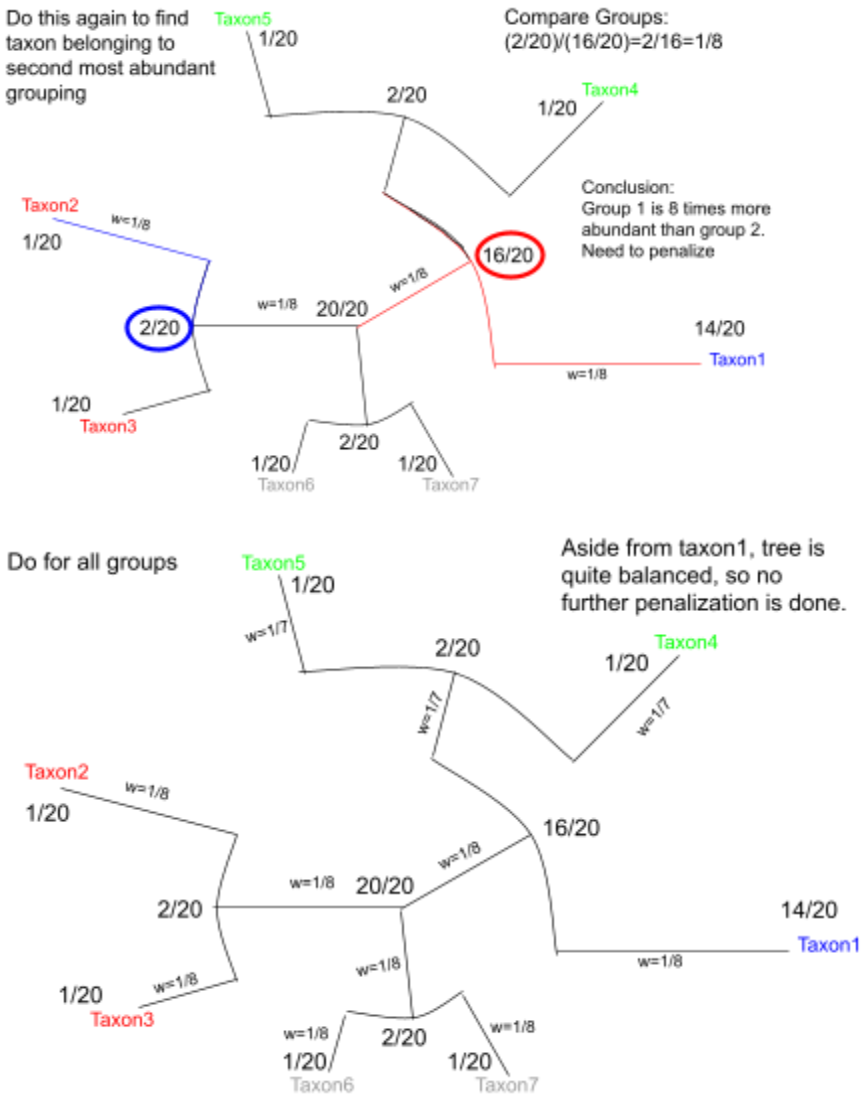


Figure 4. Gives a visual walkthrough of the finalized Dynamic Grouping Algorithm.

T1	T2	T3	T4	T5	T6	T7	S1	S2	S3	S4
1/7	1/7	1/7	1/7	1/7	1/7	1/7	1	1.0	0.1884	0.6293
3/10	1/10	0	3/10	3/10	0	0	4/7	0.8146	0.2443	0.3521
3/10	0	0	3/10	3/10	1/10	0	3/7	0.7082	0.2303	0.3402
3/10	1/40	1/40	3/10	3/10	1/40	1/40	1	1.0	0.2211	0.3460
14/20	1/20	1/20	1/20	1/20	1/20	1/20	1	1.0	0.0660	0.1345
1	0	0	0	0	0	0	1/7	0.0	0.0	0.0
0	1/4	1/4	0	0	1/4	1/4	4/7	0.3791	0.1250	0.3791
1/3	1/3	0	1/3	0	0	0	3/7	0.7628	0.3354	0.5445
1/3	1/3	0	0	0	1/3	0	3/7	0.7099	0.3121	0.7095
1/6	1/6	1/6	1/12	1/12	1/6	1/6	1	1.0	0.1987	1.0
0	1/2	0	0	1/2	0	0	2/7	0.4201	0.2771	0.4199
0	1/2	0	1/4	1/4	0	0	3/7	0.4879	0.1609	0.4878

Table 2. T=Taxon, S1=Sample Count Score, S2=Unweighted Score, S3=Initial Weighted Algorithm Score, S4=Final Weighted Algorithm Score. Unweighted classified the 5th sample as diverse, failing to see that the sample is skewed enough not to be diverse. Weighted catches this. Also note that sample 11 and 12 are differentiated properly by the update in the algorithm, with 12 being noted to be more diverse by the finalized algorithm.

Conclusion

Through our analysis of our initial weighted phylogenetic diversity algorithm, we found that the algorithm disproportionately rewarded samples with high abundances of a few species, highlighting the limitations of our algorithm and emphasizing the need for revisions to our algorithm. By introducing our finalized dynamic grouping algorithm, we addressed these shortcomings by weighting edges based on the collective abundance of descendant groups, ensuring a more balanced and unbiased species distribution. Future work would focus on refining the algorithm for larger and more complex real-world datasets, allowing for its application across ecological and evolutionary studies.

Methods

Apart from the Taxon samples examined as shown in results, data was also tested on the larger tree of hw3 species (Figure 5). Abundances for pseudo-samples were created by producing a random numbers through a normal distribution with a mean of 0 and variance of 1, then plugging the result into a ReLU function, which set half of species' abundances in each sample to 0, meaning that those species were not in the sample. The value from the normal distribution for the other half of the sample is maintained, so their abundance is random. Finally, all values in each sample are normalized, so that their abundances sum to one. Using this, 100 pseudo-samples were created. Findings were that species with the least diversity simply had less species present in their samples, and that the species with max diversity had high abundance in a part of the tree with three species that are significantly distanced from the rest of the tree. This makes sense; having high abundance of those species means that much of the sample's abundance significantly differed from other portions of the tree, since those species differ significantly from the rest of the tree per their placement on the tree.

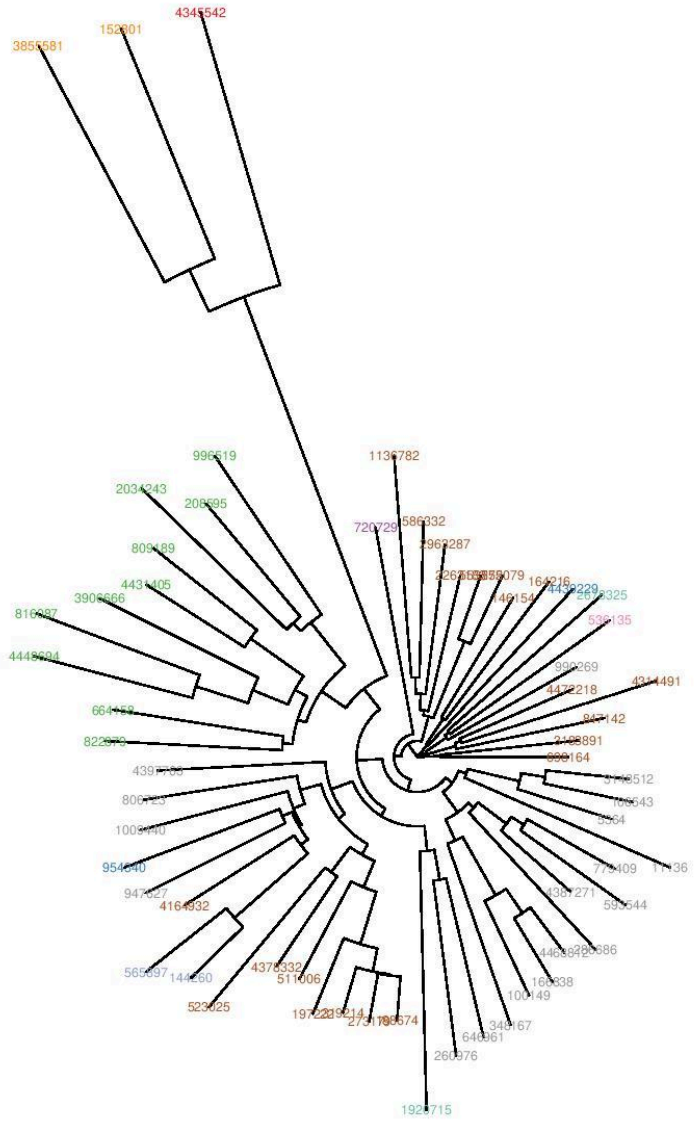


Figure 5. Shows a tree of 61 samples provided in homework 3 file data. It can be observed in the upper portion of this tree that there are three species which are greatly distanced from the rest of the tree. These are the species in which high abundance impacts diversity the greatest.

Acknowledgments

Joshua O'Neill designed, wrote, and fully debugged all five algorithms used to compute weighted phylogenetic diversity, including the finalized version of the weighted phylogenetic algorithm. Created fake samples with both HW3 Taxon data and HW3 Phylon data, tested it on all algorithms, and interpreted the results and analyzed whether these results supported or contradicted our hypotheses, and then answered the question of why that might be. As a result of this process, often made edits to the algorithms or created new ones altogether. Created all visuals for the presentation and this report, wrote and presented the approach and results slides, and wrote the results and methods section for this report.

Ethan Tenley wrote the abstract and conclusion paragraphs of this paper. He also set up the formatting of this document. Ethan also created and presented the introduction and conclusion slides for the in-class presentation. Along with all other members of the group, Ethan assisted in brainstorming ideas for the final project. Finally, Ethan set up two meetings to review the project as a group, facilitating group work.

Bang Ly Looked up previous findings and found two to three articles related to measuring the abundance in phylogeny. Bang also created a previous findings slide, presented it in class, and answered questions. Then, Bang wrote the summary of the previous findings section and citations in the final project report.

Works Cited

1. Chang, Qin, et al. “Variance Adjusted Weighted UniFrac: A Powerful Beta Diversity Measure for Comparing Communities Based on Phylogeny.” BMC Bioinformatics, vol. 12, no. 1, Apr. 2011, https://doi.org/10.1186/1471-2105-12-118.
2. McCoy, Connor O., and Frederick A. Matsen. “Abundance-weighted Phylogenetic Diversity Measures Distinguish Microbial Community States and Are Robust to Sampling Depth.” PeerJ, vol. 1, Sept. 2013, p. e157. https://doi.org/10.7717/peerj.157