Integrating transcriptomics with Startle

Kowshika Sarker
University of Illinois at Urbana-Champaign, Illinois, USA

Abstract. Recent technologies like CRISPR/Cas9 can introduce in-vivo biological systems with cells containing artificially engineered mutations at specific genomic locations. When these engineered cell(s) go through repeated cell divisions, all the descendants inherit the ancestral mutations. Information on these mutations and gene expressions captured at single-cell resolution are utilized for reconstructing single-cell level evolutionary histories, inferring cell expression trajectories, etc. Startle, and some other works, have proposed cell lineage reconstruction approaches based on the mutation information. A few recent methods have utilized transcriptomic information and CRISPR/Cas9 mutations together for cell lineage reconstruction. The phylogenies produced by Startle usually contain a high degree of polytomy. In this work, we investigate the integration of single-cell expressions with the Startle mechanism to obtain more refined cell division phylogenies and understand the expression pattern changes across the cell lineage. We formulate the problem as a maximum parsimony framework where the goal is to (1) refine the topology of an existing cell division tree and (2) assign cell expression states to ancestral nodes in the cell division phylogenies such that the difference of expression state distributions of sibling subtrees is minimized.

Keywords: Cell lineage · CRISPR/Cas9 · Transcriptomics

1 Introduction

Reconstructing single-cell level phylogeny is crucial for understanding several biological phenomena like organism development, tumor evolution, etc. Inferring cellular phylogeny based on natural somatic mutations is often tricky due to low-frequency rates and other factors. CRISPR/Cas9 and other recent lineage tracing technologies can introduce artificially engineered cells into living systems, such that these cells are very likely to obtain mutations at specific target sites of the genome, and once such a mutation is achieved, that gets inherited by all the descendant cells. Then sequencing the target sites of a cell population produces a mutation matrix, often referred to as a barcode matrix, where each cell is represented by the collection of categorical mutations obtained at target

sites. Several lineage reconstruction methods have been proposed based on barcode data - DCLEAR [1], Cassiopeia [2], Startle [9]. Different methods focus on different characteristics specific to the lineage tracing data, such as Startle introduced a star homoplasy model to account that once a CRISPR/Cas9 induced mutation is obtained in a cell, the mutation cannot get lost in any descendant. The same target site, often referred to as a character, can obtain mutations at different places of the phylogeny. Startle proposes a maximum parsimony framework to reconstruct phylogenies under the star homoplasy model.

Some studies, such as LinRace [6], and LinTIMaT [11] have paired single-cell expression with barcode profiles for lineage reconstruction. LinRace disjointly clusters cells into potential cell types, referred as cell expression states, and infers a lineage across the clusters by a trajectory inference method Slingshot [10]. Constructing a backbone tree based on the barcode mutations, the topology among cells with the same barcode profile is resolved by a maximum likelihood approach where the likelihood is dependent on the cell cluster lineage.

Broadly, this project focuses on integrating LinRace and Startle mechanisms for a prospective better lineage reconstruction approach. The following aims were presented in the proposal.

1. **Aim 1:** At the 2$^{\text{nd}}$ phase of LinRace, informing the local search with Startle topology, such as initiating the search with the Startle subtree over the subgroup of cells at consideration.
2. **Aim 2:** As Startle often produces highly polytomous trees, post-processing Startle trees by utilizing the cell state tree to refine the polytomous nodes and assigning expression state to internal nodes of the tree to understand the transition of cell types across the cell lineage.

This report is based on the 2$^{\text{nd}}$ aim. The 1$^{\text{st}}$ aim wasn't conducted.

2 Methods

On a real dataset of the *Caenorhabditis elegans (C. elegans)* species, Startle yielded cell lineages where 42.08% to 86.44% internal nodes are polytomous. The highest outdegree of a polytomous internal node varies from 9 to 214. Also, as Startle depends only on barcode data, it doesn't provide insights into how the expression states are transitioning in the cell division tree.

2.1 Preliminary

Refinement: Say, u is an internal node in a rooted phylogeny with an outdegree of d and the nodes at the other end of the d edges outgoing from u are $\{v_1, v_2, \cdots, v_d\}$. We can consider a star topology at u. A refinement of u denotes any rooted topology at u over $\{v_1, v_2, \cdots, v_d\}$ where outdegree of u is

$< d$. Before refinement, the most recent common ancestor (MRCA) of any subset of $\{v_1, v_2, \cdots, v_d\}$ is u. After refinement, there will be at least one subset of $\{v_1, v_2, \cdots, v_d\}$ for which the MRCA is a descendant of u. A refinement of a tree T is obtained by refining one or multiple polytomous nodes of T.

Our 2nd aim can be thought of as an intermediate of small and large parsimony problems - in small parsimony, we have a known tree topology and we assign states to internal nodes by optimization of an objective function, and in large parsimony, both the topology and the state assignments are inferred. In our case, we already have an initial topology produced by Startle, which we would like to partially change by refining the polytomous nodes and keeping the basic topology unchanged. Then we would like to assign expression states to ancestral nodes. We define our 2nd aim in two subproblems.
Input:

- A polytomous rooted phylogeny T with the leaves being n cells $\{cell_1, cell_2, \cdots, cell_n\}$.
- A mapping of each cell $cell_i$ to exactly one among the K clusters $\{C_1, C_2, \cdots, C_K\}$. Each cluster C_k is considered as a value of the categorical variable *expression state* of a cell.
- A rooted state tree S, where each internal or leaf node (including the root node) is labeled by one cluster among $\{C_1, C_2, \cdots, C_K\}$, and no cluster is used for labeling multiple nodes of the state tree.

Subproblem 1: The output is a refinement of T, T' such that for every node u in T that's refined in T', the difference of distributions of cell states in sibling subtrees is below a threshold. Later we define three ways of quantifying this difference.

Subproblem 2: This is a small parsimony problem over the refined tree T' where each internal node is assigned a state among $\{C_1, C_2, \cdots, C_K\}$ optimizing an objective.

2.2 Approach

Subproblem 1: We resolve the 1st subproblem by applying the UPGMA approach at each polytomous node u with outdegree $d > 2$. Let's consider a star topology over the d children $\{v_1, v_2, \cdots, v_d\}$. We perform a hierarchical clustering to refine the star topology and then insert the refined topology in the bigger tree. The approach is demonstrated in Figure 1. We visit the internal nodes of the input tree in a post-order traversal and before refining an internal node, make sure all its children are refined.

We propose the following five criteria to compute the cost for UPGMA. All five criteria produce values between 0 and 1 (both inclusive).

1. **c1: Jensen-Shannon cost:** Say, we have two nodes (can be internal or leaf) n_1 and n_2. We define the distance between the nodes (i.e., the two subtrees under the nodes) as the Jensen-Shannon distance (JSD) between

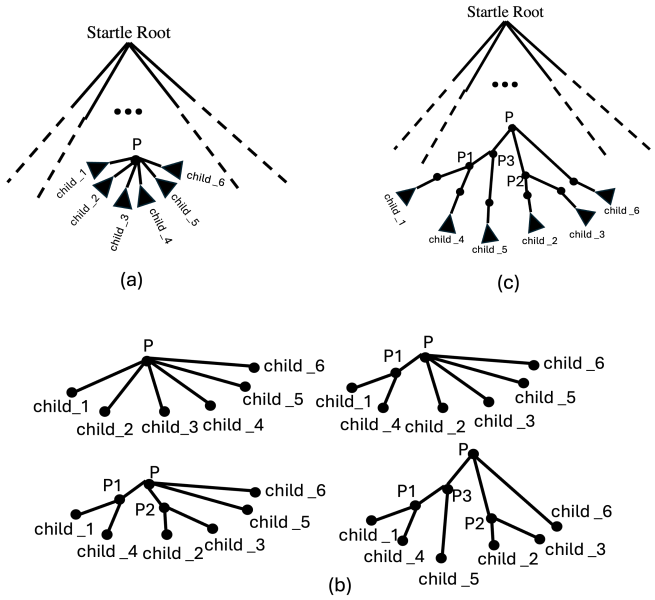


Fig. 1: Schematic of the refinement procedure. (a)Input startle tree with a polytomous node P. (b) A dummy refinement of node P. Refinement is applied on the internal nodes of the input tree in a post-traversal manner. So, no ancestor of P will be refined before P. Note that, due to using a threshold in UPGMA, our refinement is designed to stop if no pair of siblings under a polytomous node is close enough according to expression similarity. In this dummy refinement, the final refined output of P is also polytomous with 3 children. (c) Insertion of the refined node in the larger tree.

the probability distribution of the K clusters in the leaves of the two subtrees.

2. **c2: Cluster overlap cost:** Distance between two subtrees is defined as the maximum proportion of cells belonging to different clusters. Say, we have 3 clusters $\{C_1, C_2, C_3\}$. The subtree under n_1 has leaves with a proportion of $C_1 = 20\%, C_2 = 50\%, C_3 = 30\%$ and the subtree under n_2 has leaves with a proportion of $C_1 = 40\%, C_2 = 25\%, C_3 = 35\%$. Then $similarity(n_1, n_2) = similarity(n_2, n_1) = min(0.2, 0.4) + min(0.5, 0.25) + min(0.3, 0.35) = 0.2 + 0.25 + 0.3 = 0.75$. So $distance(n_1, n_2) = distance(n_2, n_1) = 1 - 0.75 = 0.25$.
3. **c3: State change cost:** If two states are the same, the cost is zero, and the cost is 1 otherwise.
4. **c4: Lineage linear cost:** Distance from an ancestor state to a descendant state is the path length between them. The cost from the ancestor to the descendant is obtained by normalizing the distance by the maximum distance in the tree. The cost of a state change where the start state is not an ancestor of the end state is infinity, we use an infinity value of 2.0 in our experiments.

5. **c5: Lineage sigmoid cost:** Distance from an ancestor state to a descendant state is the path length between them. The distance of a state pair where the start state is not an ancestor of the end state is infinity. Then we obtain the cost $c5$ from distance d as $c5 = 2((\frac{1}{1+e^{(-x)}} - 0.5))$.

Among the criteria, $c1, c2, c3$ use information only about the expression states of cells but do not use information about the evolution among the states i.e., do not utilize the state tree. The other two criteria, $c4, c5$ utilize information about state tree.

Subproblem 2: We solve this using the Sankoff algorithm and the normalized cluster path distance. We apply the three criteria c3, c4, and c5 to solve this subproblem. We cannot use the criteria c1 and c2 in the small parsimony problem of subproblem 2, because these two criteria reflect dissimilarities between distributions, not distances between states. So while these can help us to detect if two subtrees are similar enough in terms of their expression state proportions, these cannot give an assignment of which expression state to assign to a potential ancestor when two subtrees are similar enough.

3 Experimental Setup

3.1 Datasets

We use a dataset on the *C. elegans* species, for which the binary rooted ground truth cell lineage over 363 cells and single-cell expression values for 93 genes are available. Barcode profiles are not available for the cells, the ground truth lineage was constructed using cell morphology information. The barcode profiles are simulated using TedSim [7], using the ground truth tree as an input. This dataset is available at [3] and LinRace used this dataset in their benchmarking.

3.2 Experiments

In the input data, while simulating the barcode profiles, two hyperparameters of TedSim are varied - mutation rate (varied from 0.03 to 3.0) and dropout (yes/no), and 9 target sites are used. For every mutation rate and dropout combination, 10 replicates of the barcode profiles are generated for stability.
As the 3 distance measures used in the subproblem 1 produce distance between 0 and 1, 10 thresholds $\{0.1, 0.2, \cdots, 1.0\}$ are used with each distance type to generate the UPGMA refinements.

3.3 Evaluation

We use the following three metrics to benchmark our approach. All three metrics measure the extent to which the splits of the predicted trees match with the ground truth tree. A split is defined by the two disjoint leaf subsets generated by the removal of an internal edge of an unrooted phylogeny.

Say, we have two splits $S_1 = A_1|B_1$ and $S_2 = A_2|B_2$ from two trees T_1 and T_2 over leafset $\mathcal{L}$, respectively. Here, A_i and B_i the two leaf subsets of split S_i where $A_i \cap B_i = \phi$ and $A_i \cup B_i = \mathcal{L}$.

1. **Normalized Robinson–Foulds distance (RFD):** RF distance measures how many exactly same splits are shared between two trees, the normalized version normalizes the common split count with the no. of possible splits [8].
2. **Nye similarity (NYE):** This metric optimally matches branches of one tree to another and assigns a matched branch pair a score based on the size of the largest split that is consistent with both of them and then this score is normalized against the Jaccard index[5].
3. **Clustering information distance (CID):** This metric is defined based on the mutual clustering information (MCI) score [4]. The MCI score denotes that if we know which subset a leaf belongs to in one tree's split, then how likely are we to assign that leaf to the correct subset in the other tree's split? CID is the complement measure of MCI that reflects the distance between two trees.

4 Results

All results presented here are performed on the *C.elegans* dataset. The single-cell expression values for 363 cells and 93 genes are experimentally measured and the ground truth tree is manually curated by analyzing cell morphological information. However, the barcode data on mutations was not experimentally measured for this dataset. So LinRace utilizes the barcode data simulated with TedSim with varying mutation rates $(0.03, 0.25, 0.1, 0.2, 0.05, 0.3, 0.15)$ and dropout (with/without). For each experimental setting defined by a combination of mutation rate and dropout, 10 replicates are simulated.

4.1 Refinement quality with varying UPGMA thresholds

We compare the quality of the final tree with various thresholds in the UPGMA approach. The median CID of the final trees and the ground truth are presented in Figures 2. The median RFD and NYE and average RFD, CID, and NYE are shown in Figures 10, 11, 13, 12 and 14. We can see the quality of the final trees does not show a consistent trend with changing thresholds.

5 Amount of polytomy with varying UPGMA thresholds

The median no. of polytomous nodes and median of max outdegree in the refined and unrefined trees are shown in Figures 3 and 4, respectively. As expected, the amount of polytomy decreases with increasing thresholds - for each criterion, we have fewer polytomous nodes with higher thresholds and the maximum outdegree also decreases with greater thresholds. Also, compared to the unrefined trees, all criteria and threshold combinations reduce the polytomy levels by a huge margin.

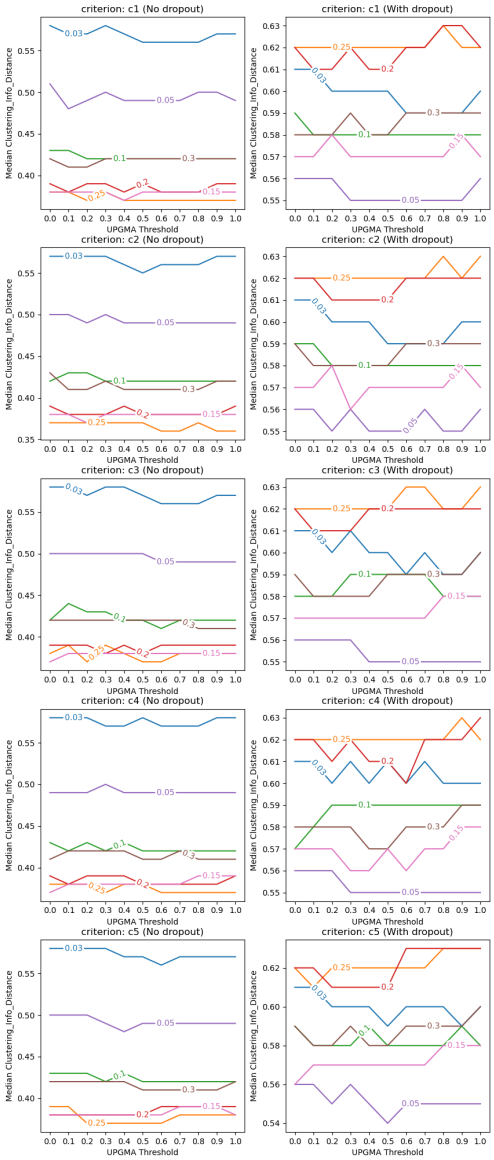


Fig. 2: Median of clustering information distance across 10 replicates of each experimental setting.

For low and medium mutation rates (0.03 and 0.15), the 5 criteria changed significantly in terms of the amount of refined polytomy. For a high mutation rate (0.3), the amount of polytomy refined by the five criteria is quite close.

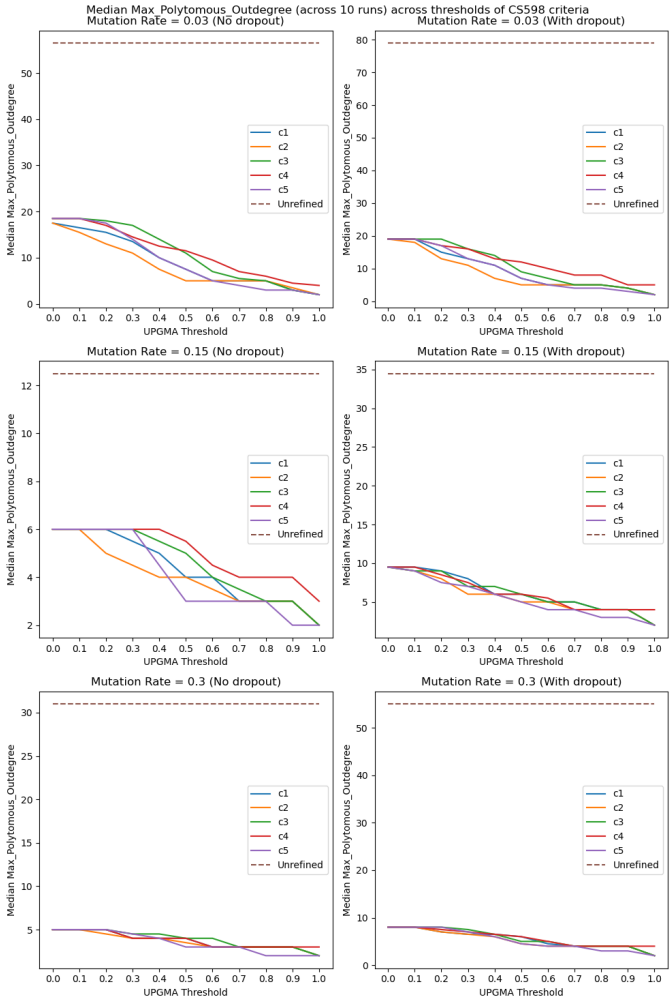


Fig. 3: Median no. of polytomous nodes among 10 replicates of each experimental setting.

We present the Spearman rank correlation of the tree qualities with polytomy level in Table 1. The correlation is taken over all the 7700 combinations of 7 mutation rates, 2 types of dropouts, 10 replicates, 5 criteria, and 11 thresholds. We can see the strong correlation of polytomy levels with tree quality - the distance measures RFD and CID are positively correlated and the similarity measure NYE is negatively correlated with both polytomy measures. This indicates refining polytomies contributes to accurate cellular lineage estimation.

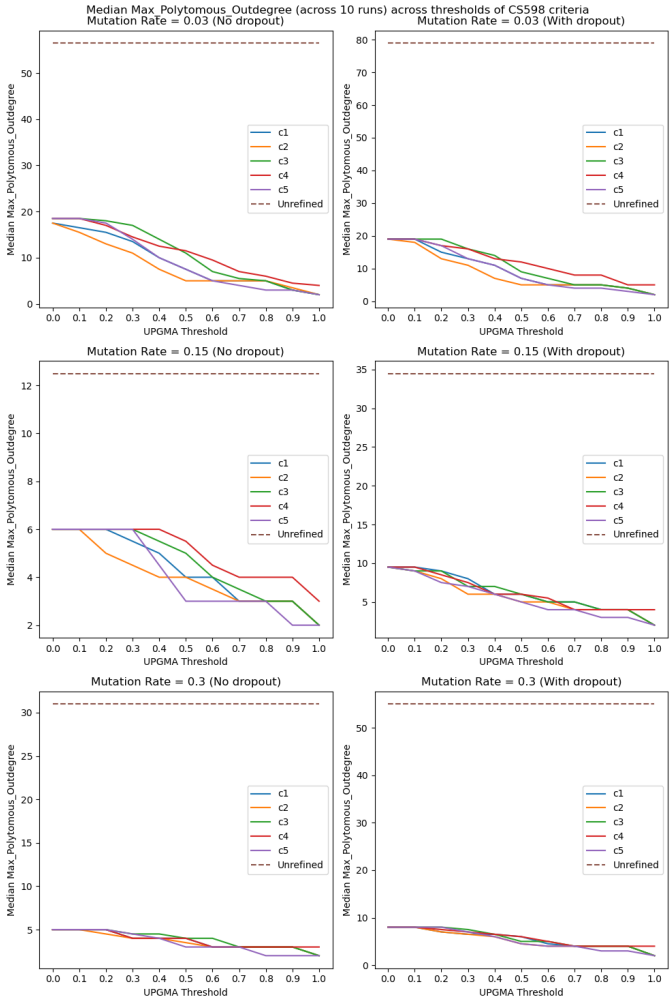
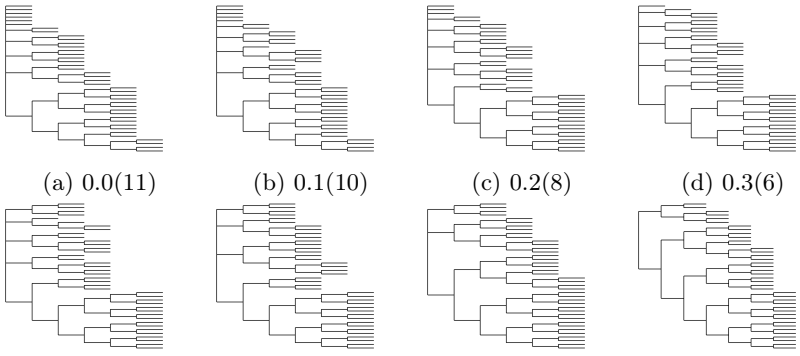


Fig. 4: Median of maximum outdegree among 10 replicates of each experimental setting.

Example of refinement We take a sample case - the 9^{th} replicate for mutation rate 0.2 with dropout - where an internal node in the Startle tree had an outdegree of 40. The refined structure of this node with varying thresholds of criterion c5 is shown in Figure 5.

	Polytomous Node Count	Max Polytomous Outdegree
Normalized RF Distance	0.131358	0.221577
Clustering Info Distance	0.216247	0.361638
Nye Similarity	-0.128474	-0.236399

Table 1: Correlation of tree quality and polytomy levels. We can see a positive correlation of polytomy levels with the distance measures RFD and CID, and a negative correlation with the similarity measure NYE. This indicates refining polytomies contributes to accurate cellular lineage estimation.



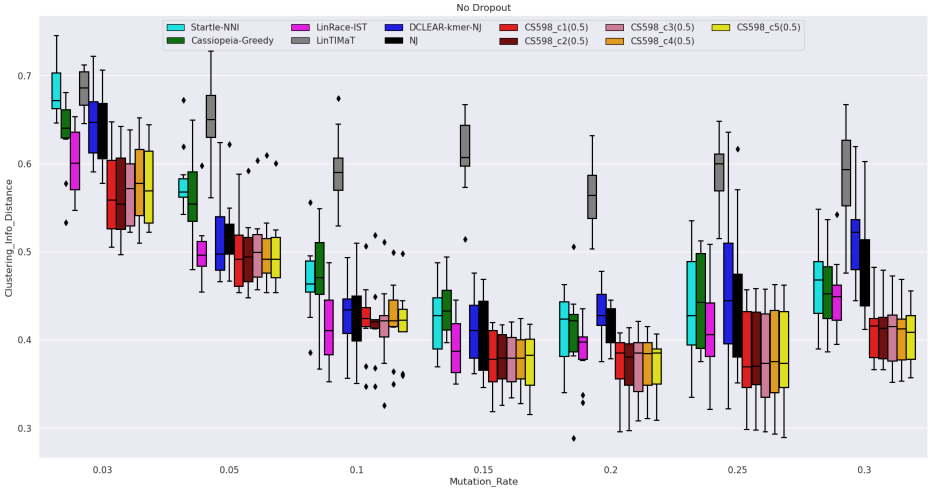
(a) 0.0(11) (b) 0.1(10) (c) 0.2(8) (d) 0.3(6)

(e) 0.4(5) (f) 0.5(4) (g) 0.6(3) (h) 0.9(2)

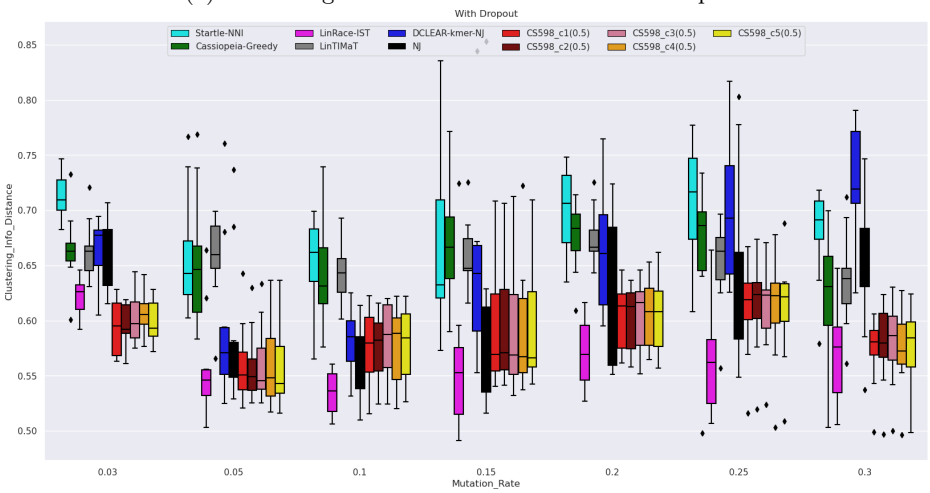
Fig. 5: Refinement of an internal node with an outdegree of 40 in the Startle tree of the replicate 9 for mutation rate 0.2 with dropout. The refinement is applied with varying thresholds of criterion c5. The numeric values T(D) denote the threshold and outdegree - such as 0.3(6) in (d) denotes that a threshold of 0.3 was applied and the outdegree after refinement is 6.

5.1 Comparison across methods

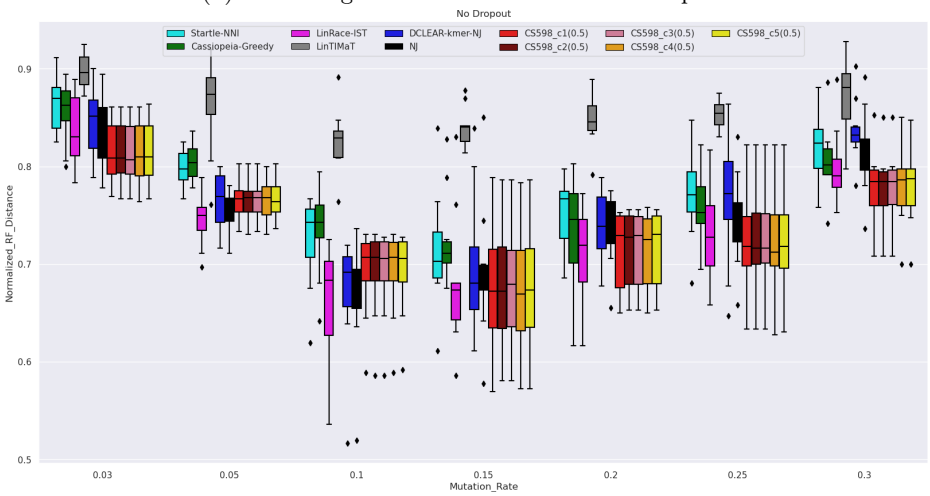
The qualities of the refined trees are compared with different methods - Cassiopeia-Greedy, LinRace-IST, LinTImaT, DCLER-kmer-NJ, and Startle-NNI in Figure 6. As seen in Section 4.1, no threshold was a clear winner for the proposed 5 UPGMA criteria. Here we pick a medium threshold of 0.5 for each criterion to compare with other methods.



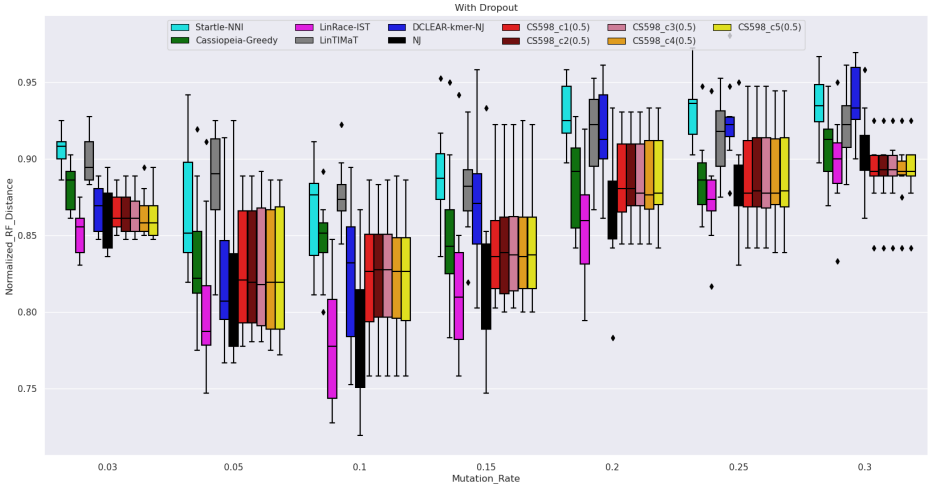
(a) Clustering information distance without dropout



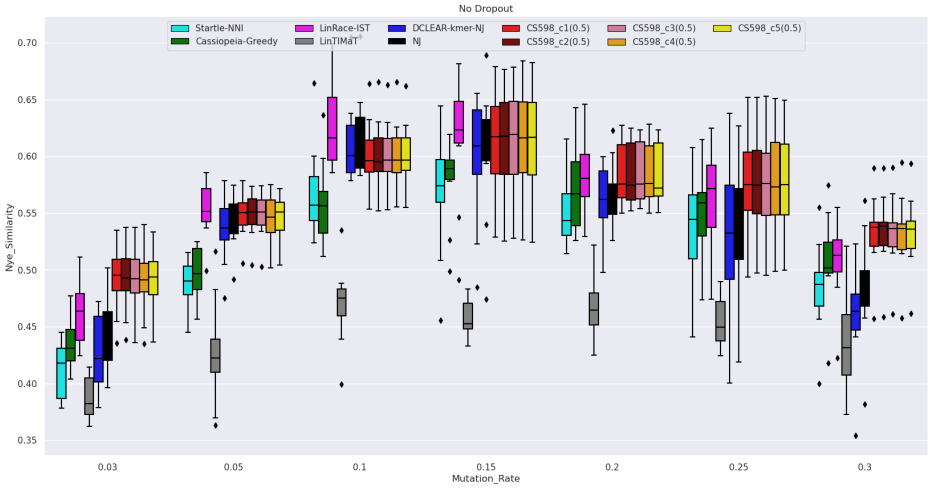
(b) Clustering information distance with dropout



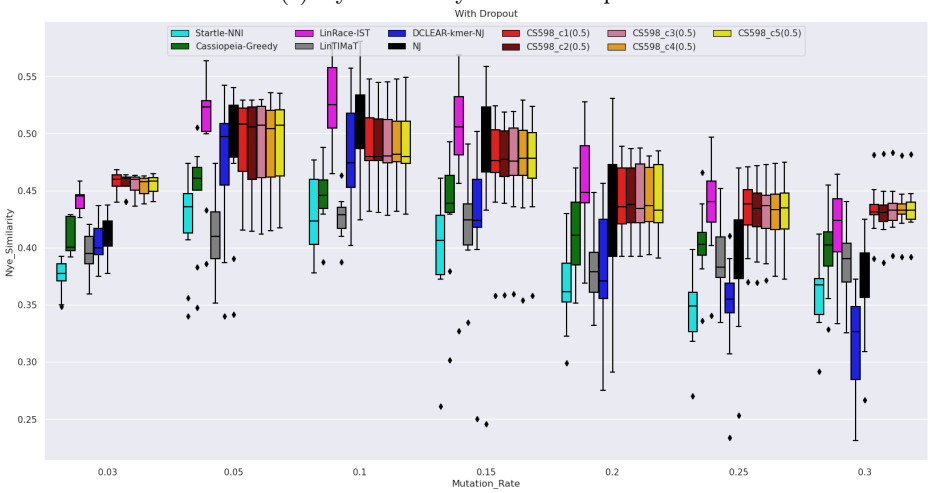
(c) Normalized RF distance without dropout



(d) Normalized RF distance with dropout



(e) Nye similarity without dropout



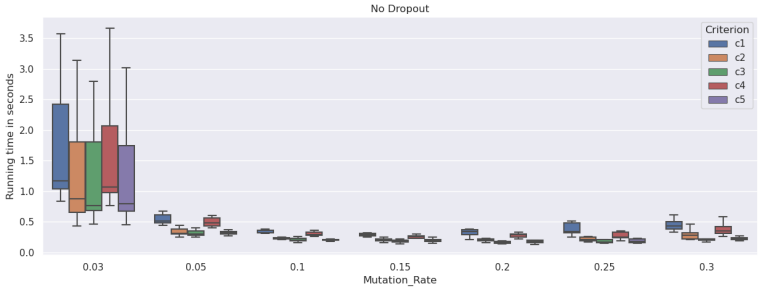
(f) Nye similarity with dropout

Fig. 6: Comparison of clustering information distance (CID), normalized RF distance (RFD), and Nye similarity (NYE) of the refined Startle trees with the trees produced by different reconstruction methods. For all the five criteria a threshold of 0.5 has been picked for this comparison. For CID and RFD higher is better and for NYE, lower is better.

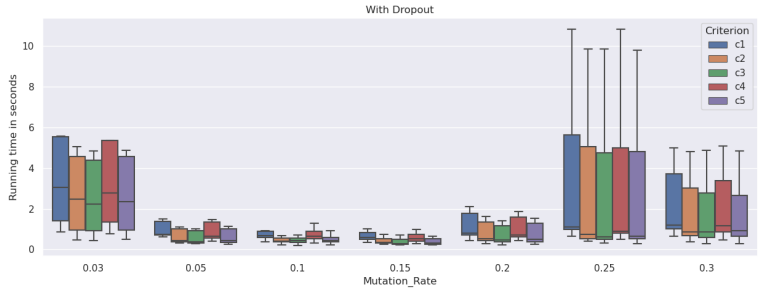
In all settings, the proposed refinements using c1-c5 improve the quality of trees compared to Startle-NNI by a huge margin, which reflects that refining polytomies informed by expression patterns can resolve cell division histories. We can see in Figure 7a, that the proposed refinements (c1-c5) yield less CID with ground truth compared to other reconstruction methods. In Figure 7d, criteria c1-c5 generate less or equal RFD compared to other methods, except for mutation rates 0.05, 0.1, and 0.15, LinRace-IST produces less RFD than our approaches. In 6e, our approaches produce better trees than other methods, except that LinRace-IST trees are better for mutation rates 0.05, 0.1 and 0.15. In Figures 7b, 6d and 6f, our approaches improve tree qualities of Startle-NNI for all cases, and do better than LinRace-IST for low mutation rate (0.03) - for other mutation rates, LinRace-IST gives better trees.

5.2 Runtime Analysis

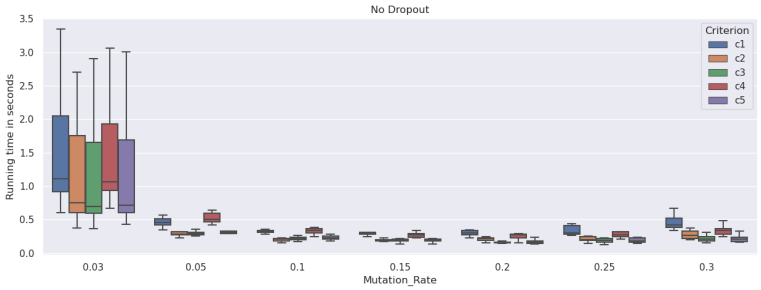
The running time of different criteria is presented in Figure 7. We can see that criteria c1 takes the longest time, c4 takes the second longest, and the other three criteria c2, c3, and c5 take almost the same level of time.



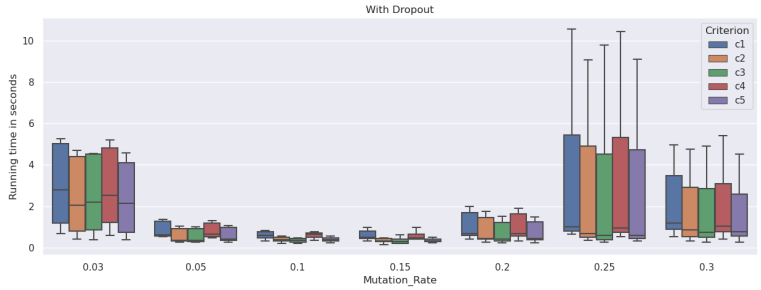
(a) Threshold = 0.0 (without dropout)



(b) Threshold = 0.0 (with dropout)



(c) Threshold = 1.0 (without dropout)

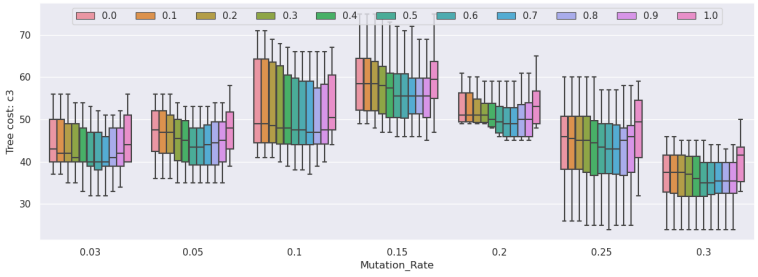


(d) Threshold = 1.0 (with dropout)

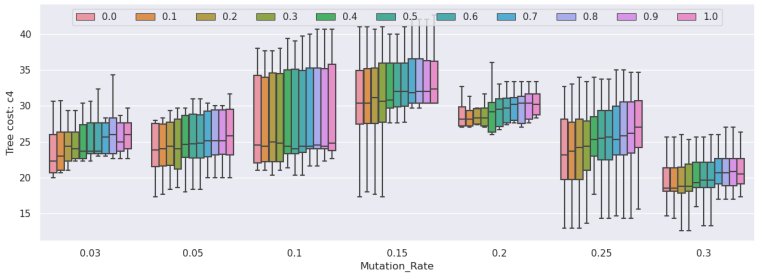
Fig. 7: Running time in seconds for refinement using different criteria. The time does not include the time taken to generate Startle-NNI trees.

5.3 Tree cost for Sankoff state assignment

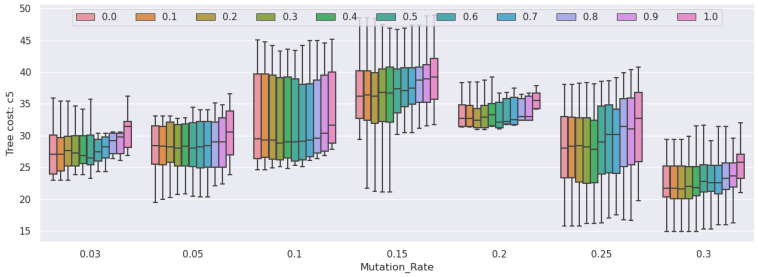
We solve the second subproblem using the Sankoff algorithm and distributions of the best tree cost with and without dropout are shown in Figures 8 and 15, respectively. When dropout exists, the c3 costs decrease with increasing mutation rates but increase again when the mutation rate gets too high. Both c4 and c5 costs increase with increasing mutation rates. Similar trends are observed in the absence of dropouts.



(a) Criterion: c3 (with dropout)



(b) Criterion: c4 (with dropout)

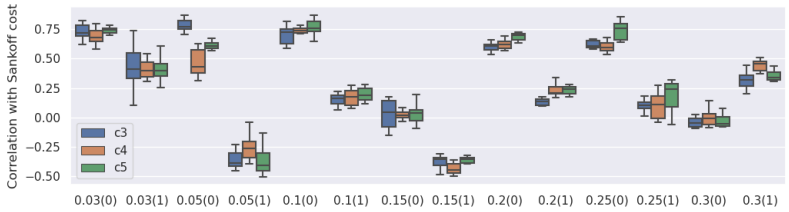


(c) Criterion: c5 (with dropout)

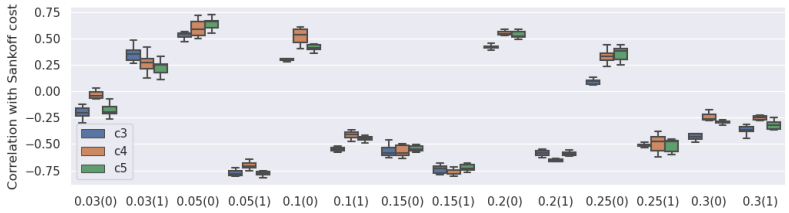
Fig. 8: Distribution of best tree cost in Sankoff algorithm for each criterion

5.4 Correlation of Sankoff cost with tree quality

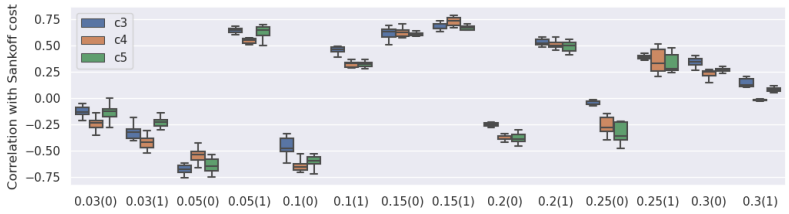
For each experimental setting (i.e., a combination of mutation rate, dropout, criterion, and threshold), we compute the correlation of the Sankoff tree cost when the state assignment is done with the same criterion. The correlation is computed with data information on 10 replicates of each setting. We show the distribution of correlations for each criterion in Figure 9.



(a) Correlation of Sankoff cost with cluster information distance



(b) Correlation of Sankoff cost with normalized RF distance



(c) Correlation of sankoff cost with Nye similarity

Fig. 9: Correlation of Sankoff cost with topology quality. For each criterion, the boxplot shows the distribution of correlation values for the 10 thresholds.

We can see, Sankoff cost shows a mixed trend of positive and negative correlations with topology quality. Among the three metrics of topology quality 9CID, RFD, and NYE), CID shows the most amount of positive correlations with the true costs of the corresponding Sankoff assignments.

6 Code Availability

Code used in this work is available at https://github.com/kowshikasarker/CS598MEB/.

7 Discussion

We proposed a parsimony approach to refine polytomous lineages reconstructed based on barcode data, using additional expression information. We are extending on a current method Startle that reconstructs lineage based on an evolutionary model accounting for homoplasy and non-reversibility of CRISPR
Cas9 mutations. We propose five criteria to compute difference across sibling pairs under a polytomous node utilizing cell expression state information. We use three criteria to solve the assignment of cell states to the internal nodes of the refined tree. We present empirical studies on real data that shows the proposed refinement strategies improve the quality of Satrtle tree topology by huge margins and beat many other existing methods.

Acknowledgments. This work is done as a project for the course CS598MEB: Computational Cancer Genomics at the University of Illinois at Urbana-Chamapign in the Spring 2024 semester. Prof. Mohammed El-Kebir instructed the course and supervised this work. We would like to thank the authors of LinRace for sharing the *C. elegans* dataset.

References

1. Gong, W., Kim, H.J., Garry, D.J., Kwak, I.Y.: Single cell lineage reconstruction using distance-based algorithms and the r package, dclear. BMC Bioinformatics **23**(1) (Mar 2022). https://doi.org/10.1186/s12859-022-04633-x, http://dx.doi.org/10.1186/s12859-022-04633-x
2. Jones, M.G., Khodaverdian, A., Quinn, J.J., Chan, M.M., Hussmann, J.A., Wang, R., Xu, C., Weissman, J.S., Yosef, N.: Inference of single-cell phylogenies from lineage tracing data using cassiopeia. Genome Biology **21**(1) (Apr 2020). https://doi.org/10.1186/s13059-020-02000-8, http://dx.doi.org/10.1186/s13059-020-02000-8
3. Liu, X., Long, F., Peng, H., Aerni, S.J., Jiang, M., Sánchez-Blanco, A., Murray, J.I., Preston, E., Mericle, B., Batzoglou, S., Myers, E.W., Kim, S.K.: Analysis of cell fate from single-cell gene expression profiles in c. elegans. Cell **139**(3), 623–633 (Oct 2009). https://doi.org/10.1016/j.cell.2009.08.044, http://dx.doi.org/10.1016/j.cell.2009.08.044
4. Nye, T.M., Liò, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics **22**(1), 117–119 (10 2005). https://doi.org/10.1093/bioinformatics/bti720, https://doi.org/10.1093/bioinformatics/bti720
5. Nye, T.M., Liò, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics **22**(1), 117–119 (Oct 2005). https://doi.org/10.1093/bioinformatics/bti720, http://dx.doi.org/10.1093/bioinformatics/bti720

6. Pan, X., Li, H., Putta, P., Zhang, X.: Linrace: cell division history reconstruction of single cells using paired lineage barcode and gene expression data. Nature Communications **14**(1) (Dec 2023). https://doi.org/10.1038/s41467-023-44173-3, http://dx.doi.org/10.1038/s41467-023-44173-3
7. Pan, X., Li, H., Zhang, X.: Tedsim: temporal dynamics simulation of single-cell rna sequencing data and cell division history. Nucleic Acids Research **50**(8), 4272–4288 (Apr 2022). https://doi.org/10.1093/nar/gkac235, http://dx.doi.org/10.1093/nar/gkac235
8. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1–2), 131–147 (Feb 1981). https://doi.org/10.1016/0025-5564(81)90043-2, http://dx.doi.org/10.1016/0025-5564(81)90043-2
9. Sashittal, P., Schmidt, H., Chan, M., Raphael, B.J.: Startle: A star homoplasy approach for crispr-cas9 lineage tracing. Cell Systems **14**(12), 1113–1121.e9 (Dec 2023). https://doi.org/10.1016/j.cels.2023.11.005, http://dx.doi.org/10.1016/j.cels.2023.11.005
10. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., Dudoit, S.: Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics **19**(1) (Jun 2018). https://doi.org/10.1186/s12864-018-4772-0, http://dx.doi.org/10.1186/s12864-018-4772-0
11. Zafar, H., Lin, C., Bar-Joseph, Z.: Single-cell lineage tracing by integrating crispr-cas9 mutations with transcriptomic data. Nature Communications **11**(1) (Jun 2020). https://doi.org/10.1038/s41467-020-16821-5, http://dx.doi.org/10.1038/s41467-020-16821-5

Supplementary Material

Refinement quality with varying UPGMA thresholds

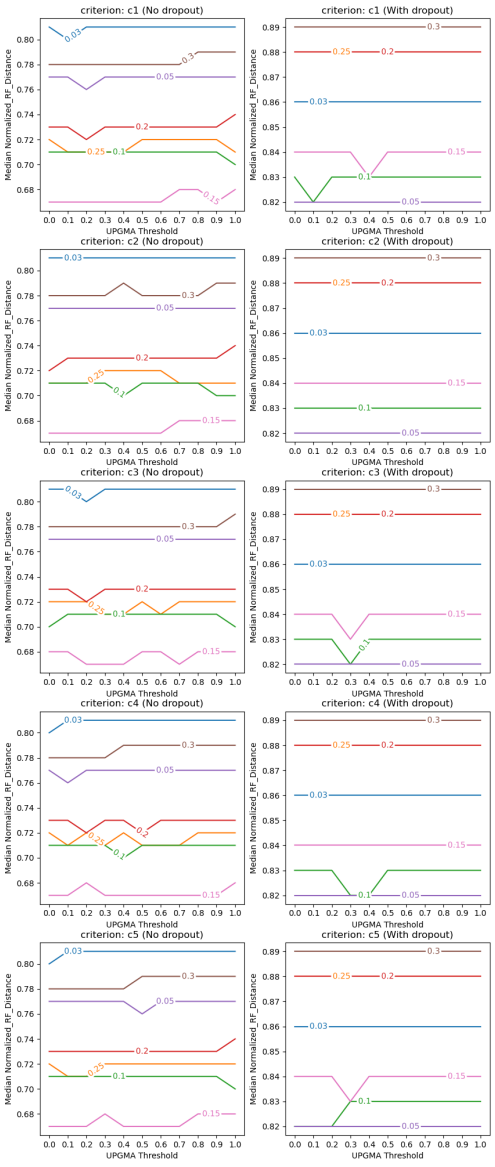


Fig. 10: Median of normalized RF distance across 10 replicates of each experimental setting.

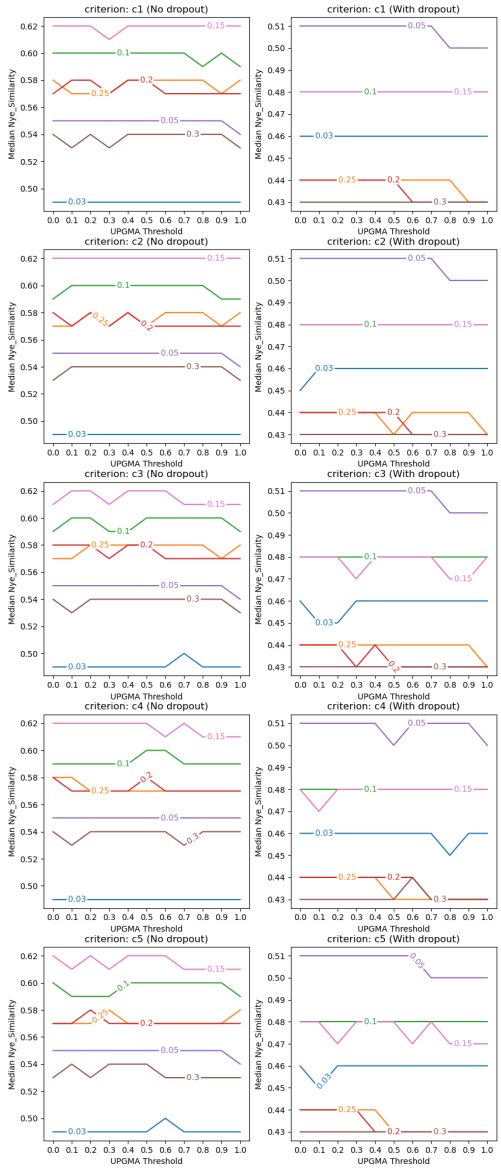


Fig. 11: Median of Nye similarity across 10 replicates of each experimental setting.

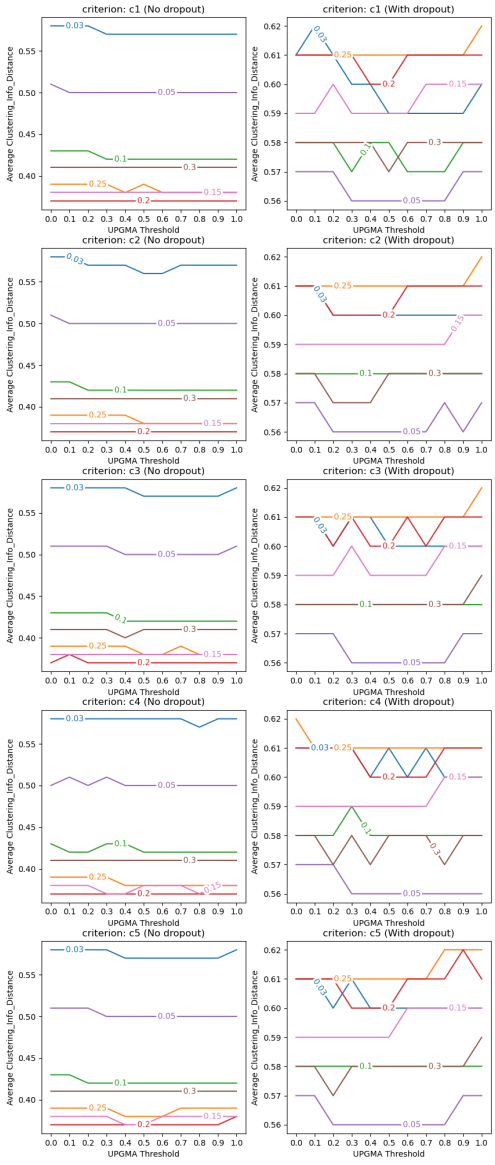


Fig. 12: Average of clustering information distance across 10 replicates of each experimental setting.

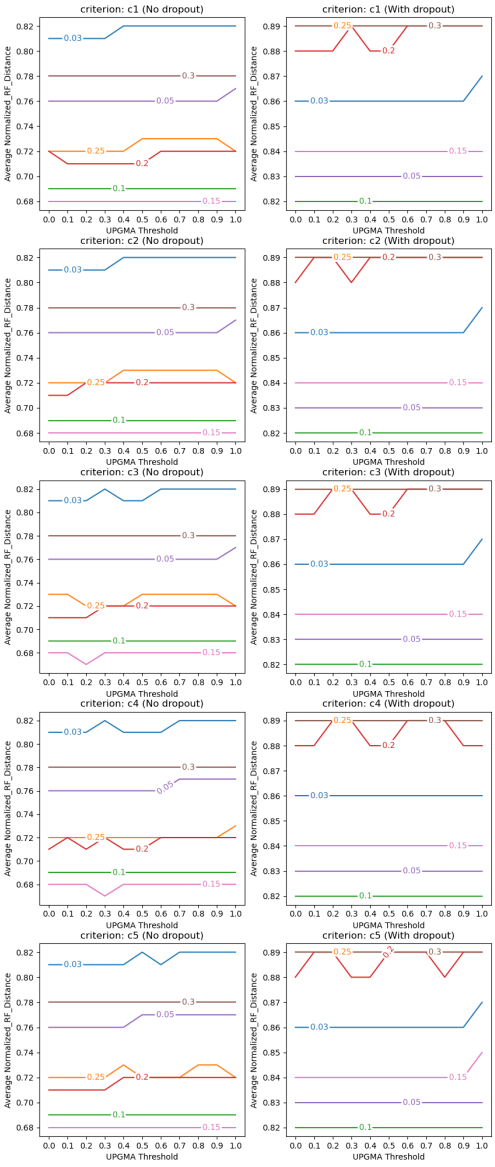


Fig. 13: Average of normalized RF distance across 10 replicates of each experimental setting.

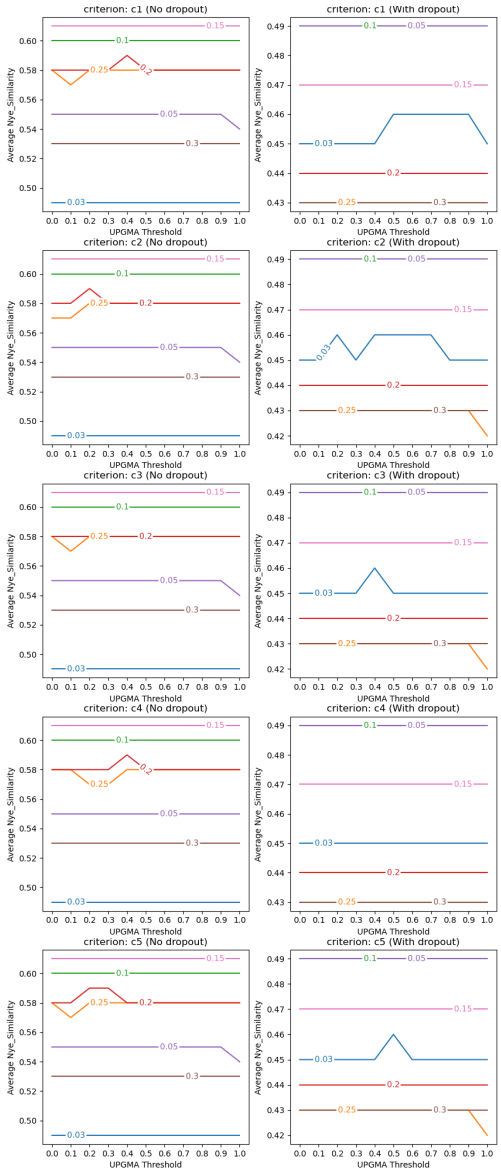
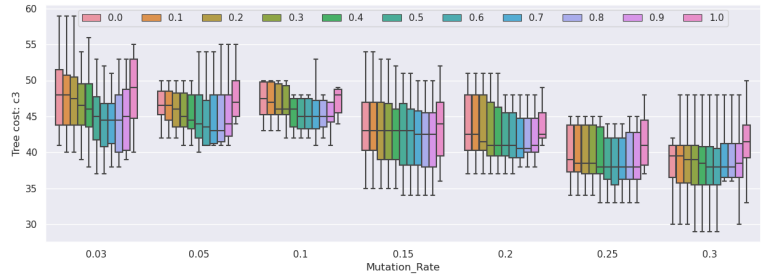
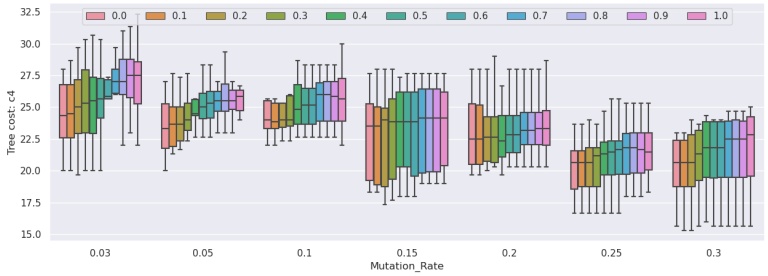


Fig. 14: Average of Nye similarity across 10 replicates of each experimental setting.

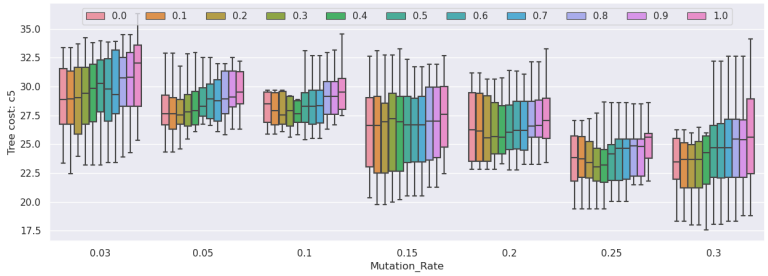
Sankoff cost distribution



(a) Criterion: c3 (without dropout)



(b) Criterion: c4 (without dropout)



(c) Criterion: c5 (without dropout)

Fig. 15: Distribution of best tree cost in Sankoff algorithm for each criterion