**Single-Cell Sequencing Phylogeny: Validation Report**

**I. Methods**

**A. Phylogeny Generation**

We used three independent distance-calculating methods to cluster our data set: L1 distance, L2 distance, and CNT distance. Our program clustered the data set once by L1 distances, L2 distances, and CNT distances separately using a neighbor-joining function.

**i. L1 Distance Clustering:**

The L1 distance measure between two Copy Number profiles is the sum of the absolute values of each positional difference between each profile, along all positions. L1 distance analysis seeks to cluster groups of samples where each L1 distance between one sample and another is minimal.

ii. L2 Distance Clustering:

The L2 distance between two profiles is the sum of the squares of each positional difference. L2 distance analysis seeks to minimize L2 distances.

iii. CNT Distance Clustering:

The CNT (Copy Number Transformation) distance, which is best comparable to edit distance, is a method proposed by Ron Shamir, Meirav Zehavi, and Ron Zeira. The CNT distance between two profiles is the minimum number of transformation operations required to transform one profile to the other. Transformation operations consist of single-copy deletions or amplifications along a contiguous segment of a Copy Number profile.

**B. Statistical Validation**

We expect each cell sample to cluster into six main clusters. The justification for this is that all samples were taken from six total distinct sites, and it is reasonable to expect that cells from the same site cluster under the same subtree.

We also expect each cluster to show statistically significant ancestry according to evolutionary distance within each cluster. For instance, two profiles occurring within cluster GBM07-SP2 should have a smaller distance than a profile occurring in GBM07-SP2 against a profile occurring in GBM07-SP3.

Using the above assumptions, we use a t-test analyzing the sum of pairwise distances of our tree against a mean of sums of pairwise distances from several randomized trees to validate the accuracy of each clustering method. For every test, we would do the following:

1) Extract all ‘leaves’ or cell samples from an input phylogeny (L1 / L2 / CNT).

2) Separate and cluster the leaves together by the region from which each leaf was sampled (six total regions).

3) Sum the pairwise distances between leaves within each cluster. Take the sum of all six cluster-sums and use this as our test distance.

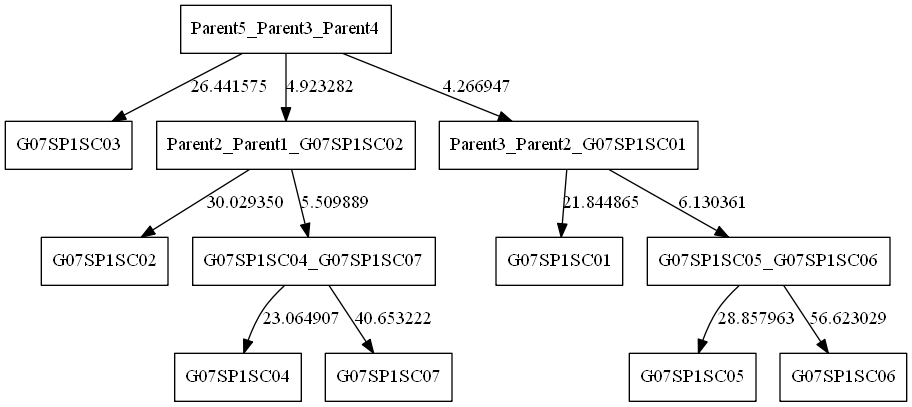
*Note:* The sum we are calculating here is different from simply taking pairwise distances between all possible leaves, even though these distances exist within the tree’s data structure. For instance, we do not calculate distance between a leaf in cluster GBM07-SP1 and a leaf in cluster GBM07-SP3 because we want to assess the accuracy of how our phylogeny clusters leaves into groups of samples.

4) For each trial, randomize the order of all the leaves and re-cluster them into six randomized groups. Re-calculate pairwise distances within each cluster again. Do this for every trial and take the mean of all of these distance sums. Compute the standard deviations for all of these sums as well.

5) Compute a t-statistic with degrees of freedom corresponding to the number of *n* trials subtracted by 1. Compute a p-value using this t-statistic comparing the according to our input tree to the mean distance sum of the *n* randomized trees.

**II. Analysis and Results**

Figure 1. Sample Phylogeny Structure with Six Cell Sample Profiles



*Note:* For all internal nodes, the labelling scheme is as follows:

[Current Node Name] \_ [Parent Child 1] \_ [Parent Child 2]

Where child nodes that are not themselves parents (i.e. leaf nodes) are the only nodes listed within the name. **The numbering of each parent only serves to distinguish groups of merged samples, and is done mainly so that the overall phylogeny can be legible.**

*Note:* All phylogenies are attached in the folder containing this document.

Figure 2. A Comparative Analysis of T-statistics and P-values for each Clustering Method

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Trials** | **T-statistic** | **P-value** |
| L1 – Distance | 10 | 17.242 | 1.672e-08 |
| L2 – Distance | 10 | 14.435 | 7.866e-08 |
| CNT – Distance | 10 | 16.211 | 2.867e-08 |

**III. Commentary**

Overall, it seems that each method clusters the data set accurately, to the extent that the clustering for each sample occurs much more significantly than by chance.

Each method saw a p-value of nearly 0.0 with all trials. Assuming that our test is a fair measure of accuracy, this would show that each method independently clusters the data set significantly, but looking at the p-values does not help to show which method does so the best. To this end, comparing the t-statistics directly may be a better strategy.

Tyler’s initial hypothesis that L1 distances gave to more overall nonrandom clustering when compared to L2 distances was vindicated by these results; additionally, we found that our implementation of the CNT algorithm gave worse clustering results than the L1 distance method. There could be many possible sources of error contributing to the differences in P-value (possibly due to the method of validation itself), and running tests with more trials is computationally very expensive, but with more time, we can validate each algorithm more accurately.

We have verified the correctness of all of these methods on multiple test cases, so each distance-calculating function should behave as expected.

**IV. Citations**

Shamir, Ron, Zehavi, Meirav, and Zeira, Ron. “A Linear-Time Algorithm for the Copy Number Transformation Problem”. *School of Computer Science, Tel Aviv University (2016).*