**Single-Cell Sequencing Phylogeny: Fall 2017 Report**

**I. Rationale**

It is nowadays widely accepted that cancer is an evolutionary disease. However, the fundamental nature of evolutionary adaptation within cancer cells is still largely unknown. Consequently, analysts have tried to apply evolutionary models to genomic variation within single tumor cells to hopefully gain a better understanding of the origins and behaviors of tumor cells in general.

One particularly interesting topic in this study is the problem of generating phylogenies for cancer cells. Under the general assumption that we can treat individual tumor cells as if they are organisms from individual species, generating phylogenies is believed to provide insight into the genetic lineage of tumor cells as they proliferate in a patient’s body. The key question is how this can be done efficiently and accurately, which this project attempts to explore extensively. The hope is that, by implementing and comparing multiple established algorithms of phylogeny generation, we are able to combine some of them into a novel method that best handles large-sample, large-genome data sets of individual tumor cell genomes.

**II. Approach**

The project addresses this question by analyzing genetic marker data from FISH (fluorescent in-situ hybridization) of several tumor cells from two patients. Data sets consisted of multiple *copy number profiles,* which are arrays of copy numbers of specific genetic markers found throughout a tumor cell’s genome via FISH. To address the main research question, algorithms in this project attempt to construct phylogenies from single-cell sequencing data sets by directly calculating evolutionary distances between copy number profiles and constructing minimum-distance phylogenies from them.

Work on the project began as a continuation of a past student’s implementations of L1 (absolute value) and L2 (Euclidean) distance measures, with a neighbor-joining phylogenetic algorithm. A new distance measure, CNTP, was implemented over the summer, and several statistical validation and comparison tests over these three methods were done this semester.

Validation tests included permutation tests over trees constructed from each of the three distance measures, with modifications such as simple moving-average smoothing over multiple fixed intervals of copy number profiles and subsample testing for groups of profiles.

Research also began on two new algorithms, the ASTRAL species phylogeny method, and FISHtrees, a previously implemented algorithm. Current efforts are to combine these two to implement yet another method for comparison.

**III. Results**

A key hypothesis at the onset of this research was that validation performance should improve as the distance measure and / or phylogeny algorithm better mimics actual biological processes in genomic variation, ex. copy number deletion & amplification. According to this, the CNTP method, which implicitly accounts for deletions and / or duplications along a contiguous segment of genetic markers, should perform the best in validation tests, with L1 and L2 (respectively) behind it.

Given the above hypothesis, the initial results we found were unexpected. Performance, from best to worst respectively, ranked as L1, CNTP, L2, with L1 and CNTP exhibiting similar P values upon validation. The best we can say based on this is that L1 and CNTP exhibit similar performance, with L2 exhibiting much worse performance.

The most efficient method by far is L1, but this is to be expected: CNTP utilizes dynamic programming to calculate edit distances between profiles, whereas L1 directly calculates pairwise absolute-value differences between profiles. Modifications to increase time / memory performance can still be made in the future.

**IV. Setbacks**

Unfortunately, because no successful effort was made to empirically analyze the complexity of each algorithm and improve memory / time performance accordingly, all methods take a significant amount of time to generate their respective phylogenies. All attempts ended in fatal execution errors and / or incorrect outputs, so this problem is left for further exploration for later on in the project.

Furthermore, while permutation tests provide a fast and simple way of assessing accuracy in phylogeny construction, they are not the best method for such a task. Consequently, future work may concern improving and / or adding methods for validating each distance measure effectively.

Finally, the ASTRAL / FISHtrees algorithm was not fully implemented due to time constraints. Future work will focus on completing this algorithm and running validation tests to assess performance on the bulk data set.

**IV. Conclusions**

As of now, much of the analysis is inconclusive with regards to overall performance. L1 continually outperforms both L2 and CNTP for the bulk data set, however CNTP outperforms both for smaller sample sets of fewer than ~100 profiles. Smoothing improves CNTP’s performance somewhat, but improves the other methods’ performances by equal ratios for the bulk data set.

The results of the ASTRAL / FISHtrees inclusion are also inconclusive, since an algorithm that encompasses both has not been fully implemented yet.

**VI. Future Directions**

Multiple opportunities exist for expanding this project to include other methods of distance measure and phylogeny generation. For one thing, current work on implementing ASTRAL / FISHtrees could be continued; furthermore, extensive research on validating such methods remains to be done. Finally, the implications of current results have still yet to be fully determined, as much work into the relation between phylogenetic information and clinical information regarding cancer is also needed.