**Project Name: A Novel Algorithm to Infer Cell Lineage Tree from Enhancer Profiles**

1. **Background and Significance**

ChIP-seq is capable of annotating binding sites of DNA-associated proteins by combining chromatin immunoprecipitation and DNA sequencing (Johnson Et al, 2007). In addition, ChIP-seq can be used to characterize histone modifications, which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation (Bannister & Kouzarides, 2011). Genome-wide profiling of certain cis-acting DNA elements is therefore possible since histone modifications are well known indicators of these features. Promotors are identified as proximal regions marked by H3K4me3, meaning that the Lysine 4 location on histone H3 protein is tri-methylated, and enhancers as distal regions marked by H3K4me1, which can be sub-categorized as active or poised based on the presence or absence of H3K27ac (Calo & Wysocka, 2013). Profiling of enhancer sites (H3K4me1 signal) across 17 cell types in the hematopoietic cell lineage (Figure 1) revealed unique enhancer landscapes for each cell type, which suggests that enhancers are gained or lost as cells differentiate. Indeed, when comparing enhancer profiles of hematopoietic stem cells and terminally differentiated cell types like monocytes and T-cells, 90% of ~50000 identified enhancers changed state. It was also observed that closely related cell types possess more similar enhancer patterns compared to those that are further apart (Lara-Astiaso Et al, 2014).

From the information above, we inferred that a cell lineage tree can be modelled using maximum parsimony, a concept borrowed from studies of the phylogeny of different species. In evolution, maximum parsimony assumes that the phylogenetic tree requiring the fewest mutations (usually DNA base changes) is most likely to be biologically correct. Adhering to maximum parsimony, we hypothesized that the correct/optimal cell lineage tree will have the smallest number of total enhancer state changes. Using the hematopoietic cell lineage as the training dataset, the aim of this project is to devise an algorithm that reads enhancer profiles of multiple cell types as input, computes the “distance” between each cell type pair, and searches for the optimal cell lineage tree that would minimize total enhancer distance. The completed algorithm and software should provide easy and efficient modelling of cell type relationships for any given group of cell types, which can be a handy tool for researchers from multiple disciplines.

1. **Innovation**

Characterizing and distinguishing between different cell types in a lineage has mostly relied on identifying cell surface markers in the past. For example, CD4 is specific to Helper T cell, CD20 to B cells, and CD34 can only be found on progenitor cells. Using enhancer profiles for this purpose is a pretty novel method, having only begun around two years ago, but validation experiments have shown it to be much more precise and accurate than using cell surface markers. Also, automating the inference of cell lineages using maximum parsimony and basing it solely on enhancer profiles have never been done before in the field to our knowledge.

1. **Methodology**

To establish a distance metric for quantification of enhancer differences between the cell types, Pearson correlation values (r) of H3K4me1 (enhancer) peak intensities were computed for all possible pairs of cell types to obtain a correlation matrix. The correlation matrix was then converted to the distance score matrix (1-r) (Figure 2). This was based on the assumption that if two cell types are closely related, they should have low number of enhancer state changes and therefore high correlation in the enhancer peak intensities. An overall score for any given tree in the adjacency matrix format, which was chosen to represent a tree computationally, can be obtained by looking up the distance score matrix. Low scores suggest that the tree is most likely to be correct under the maximum parsimony hypothesis.

The ideal way to infer the optimal cell lineage tree would be to find all possible labelled binary trees, score them, and return the one with the lowest score. However, this proved to be infeasible since the number of possible trees grows extremely fast as more cell types are added into consideration: from 30 with 3 cell types to 4.6113022e+22 trees with 17 cell types. Therefore, we decided to use a more heuristic, Monte Carlo-like approach. We simulated the building of the binary cell lineage tree by developing a probabilistic model of edge adding from the correlation matrix. In this model, given a node (cell type) as parent, the cell types that are highly correlated with the node would have higher probabilities of being added as the child, which is consistent with the assumption that two closely related cell types should have highly correlated enhancer profiles. To explore more tree structures and to take into account that some cell types might differentiate into only one cell type or none at all, a probability of not adding a child to the parent (null probability) was incorporated into the model. However, to avoid the generation of incomplete trees, the null probability was removed when adding children at the last node of a given layer of the tree if no nodes were previously added on that layer.

The simulation can be iterated many times to create a distribution of potential cell lineage trees, which is a parameter that the users can decide. The trees with the lowest scores then can be obtained and visualized, which was done using the data.tree package in R.

1. **Results**

Five cell types from the hematopoietic cell lineage: common myeloid progenitor (CMP), granulocyte-macrophage progenitor (GMP), megakaryocyte-erythroid progenitor (MEP), multipotent progenitor (MPP), and erythrocyte A (EryA), were chosen to form the pilot dataset. Results after 5000 simulations showed that the expected/validated cell lineage tree was generated with the second highest frequency, appearing 184 times (3.68 %) with the distance score of 1.98. However, the tree that was generated the most, appearing 198 times out of 5000 (3.98%), had an identical distance score of 1.98 (Figure 3). Further observations revealed that the two trees possess the same combination of cell-cell edges only with different arrangements. The histogram of the distance scores for the 5000 simulated trees showed that 1.98 is the lowest score generated, which is what was expected from the algorithm. The distribution of the histogram further supported the generation of lower scoring trees (Figure 4). The total run time of this algorithm using the five cell type dataset is around 15.5 seconds.

1. **Conclusions**

The results from the five cell type dataset indicates that our algorithm favors the generation of low scoring trees. Future directions include improving the root selecting process and null probability modelling, integration with other metrics (e.g. Spearman, Euclidean) in addition to Pearson correlation, incorporation of more cell types from the hematopoietic cell lineage into the dataset, utilization of other datasets not from hematopoietic cell lineage to test the robustness of the algorithm, and the development of a user interface.

1. **Figures**

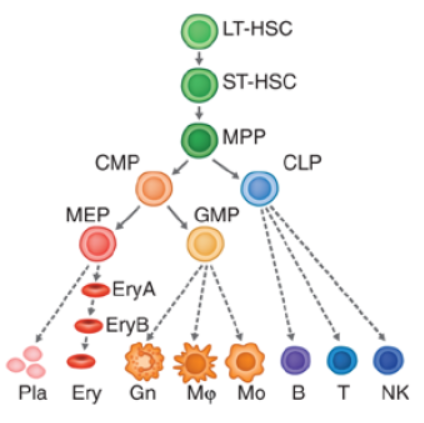
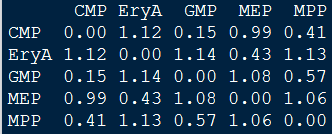
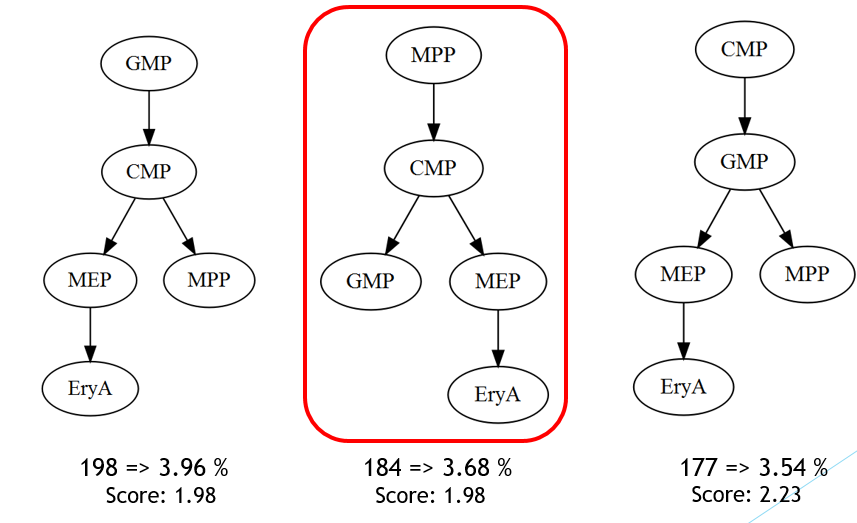


Figure 1. Hematopoietic cell lineage. Retrieved from Lara-Astiaso Et al., 2014

Figure 2. Distance score matrix computed from the five cell type dataset.

Figure 3. The top three trees generated using the five cell type dataset, with the expected/correct tree circled in red. The frequencies and the distance scores are recorded in the bottom.

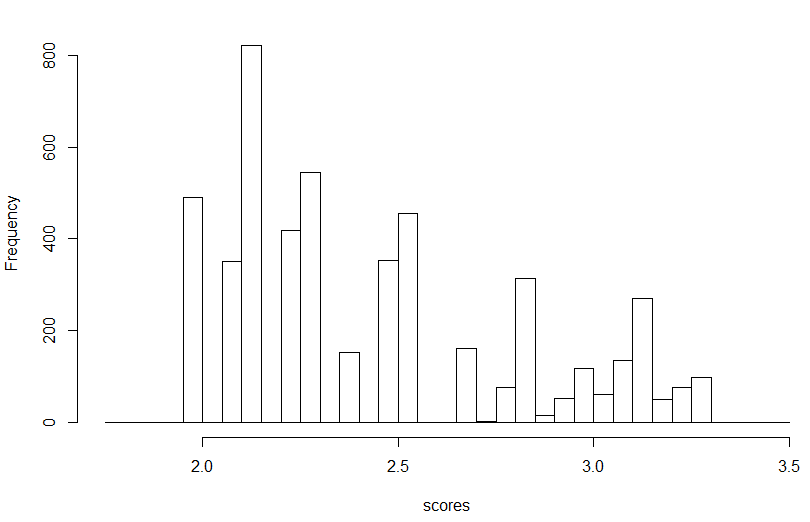


Figure 4. The histogram of the distance scores for the 5000 simulated trees.

1. **References**

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research, 21(3), 381–395. http://doi.org/10.1038/cr.2011.22

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how and why? Molecular Cell, 49(5), 10.1016/j.molcel.2013.01.038. <http://doi.org/10.1016/j.molcel.2013.01.038>

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., …Amit, I. (2014). Chromatin state dynamics during blood formation. Science, 345(6199), 943–9. <http://doi.org/10.1126/science.1256271>

Johnson DS, Mortazavi A et al. (2007) Genome-wide mapping of in vivo protein–DNA interactions. Science 316: 1497–1502