ABSTRACT: Stem cell fate differentiation is a complex biological process with important applications. Among other use cases, certain types of multipotent stem cells play a critical role in cancer treatment. Advancements in cellular assays over the past decade have provided researchers with comprehensive datasets for single-cell gene expression analysis, where one can isolate specific cells in a heterogeneous set of stem cells and elucidate genetic factors responsible for varying states of differentiation. Using network models, we attempt to show optimal fate differentiation paths for stem cells by modeling each cell as a state to predict fate progression solely through genetic markers. We successfully implemented a uniform cost search and a random walk over the Erdos Renyi and Kleinberg network models, finding sophisticated network-based models for mapping cell differentiation.

1 INTRODUCTION

1.1 Background

Early stem cell fate differentiation is one of the most important and challenging questions in modern biology. Among other applications, certain types of multipotent stem cells play a critical role in cancer treatment. Hematopoietic Stem Cells (HSCs) are a critical multipotent cell type that have the ability to differentiate into both lymphoid and myeloid blood cells and are one of the most promising candidates for clinical stem cell therapy today.\(^1\)

There are a variety of factors that influence early stem cell fate including both environmental factors and intracellular components, particularly gene transcription factors.\(^2\) The scientific community currently uses bulk assays to learn more about cell population characteristics. Biological assays are unfortunately both expensive and time consuming. By definition, assays also analyze cells in bulk. One issue with this approach is that the bulk cells being analyzed are at different stages in their differentiation process, which means that important characteristics about cells in various stages of the differentiation process can be lost in global averages.

For the reasons outlined above, it would be ideal to have an alternate analysis strategy to biology assays which would enable us to study cell fate differentiation at an individual cell level.\(^3\) Our aim in this project is to apply network models to the problem of stem cell fate declaration to infer candidate regulatory networks from gene expression data with some level of accuracy.
We will start with a single-cell gene expression dataset of mouse HSCs that was collected and studied by Moignard et al. The original dataset used by Moignard et al contains 3935 mouse embryonic stem cells belonging to five different cell classifications: 624 cells of type PS, 552 NP, 1005 HF, 983 4G, and 770 4GF (figure 1). These classifications represent broad cell types corresponding to stages of stem cell maturation. The normalized expression levels of 46 genes are also reported for each stem cell in the dataset.

Moignard et al. used Boolean networks, among a number of unsupervised statistical learning methods, to cluster the HSCs and identify which gene transcription factors control differentiation. For our analysis we will attempt to expand on their work using two different network models, the Erdos Renyi and Kleinberg models. There is currently not a biological assay that enables scientists to track gene expression changes of the course of cellular differentiation. Our task is to model this development using these two network models as a path an undifferentiated or primitive stem cell takes to its final, differentiated state.

2 APPROACH TO PROBLEM

We approach this problem by using two different paradigms for assigning terminal types given primitive cells: Uniform Cost Search and a random walk process. We evaluate each paradigm over both the Erdos Renyi and Kleinberg network models. For each of these models, we use the same framework for state definition.

Because no temporal assay exists for tracking an individual cell’s genetic evolution throughout fate differentiation, we instead model each of the original cells in our dataset as a state into which a cell can enter at a point in time during its differentiation. We consider each cell in the dataset, uniquely characterized by its gene expression levels and embryonic state, as a node in a network or possible state in a state-based process. As a result, every network model we test involves the same set of nodes or vertices the 3935 cells in the Moignard et al. dataset. States have the following form:

\[ S_i = ([g_{i1}, g_{i2}, \ldots, g_{ij}, \ldots, g_{i46}], X_i) \]

for \( i = 1, 2, \ldots, 3935 \)

\( g_{ij} \) is the expression level of the \( j^{th} \) gene in the \( i^{th} \) cell, \( X_i \in \{PS, NP, HF, 4G, 4GF\} \). We will now demonstrate how these states are used in our two implementations.

2.1 Uniform Cost Search Model

The first paradigm we use to assign terminal types given primitive cells is the Uniform Cost Search algorithm (UCS). With this paradigm, we build a graph using a given graph model, then model cell differentiation as a search problem and use UCS to find shortest paths to differentiated fates for each progenitor cell.

Through search, we deterministically explore every single path from a progenitor cell to a terminal fate, which involves either an erythroid or endothelial cell fate. Although this approach is simpler from a modeling perspective, it is computationally intensive because UCS samples the entire state space for every progenitor cell to find its shortest path. Additionally, the search problem model assumes differentiation is deterministic.

Start State: Every possible progenitor cell (PS, NP, HF).

Successors/Actions:

Baseline Model - We construct a baseline graph model by building a graph where every cell state has an edge with every possible successor cell state. Below are the mappings of current states to accessible successors.

- PS → NP
- NP → HF
- HF → 4G, 4GF

Erdos Renyi Model - The Erdos Renyi model is a model for generating random graphs. We construct the Erdos Renyi graph model by assigning an edge between a cell state and a possible successor state with probability \( p = 0.1 \).

Kleinberg Model - The Kleinberg model is a graph generation model that produces graphs with small diameters and high clustering. We are optimistic about the possibility of using the Kleinberg model as both of these properties could serve us well in predicting genetic regulatory networks. Substantial local clustering is a promising characteristic in a predictive network as a cell is more likely to stochastically change its gene expression profile within a single time-step to a very similar profile than one farther away. We seek a network model with a limited diameter as a primitive cell must be able to transition with a limited number of time-steps into some terminal cell state.

We construct the Kleinberg graph model by assigning an edge between a cell state \( v \) and a possible successor state \( w \) with probability proportional to \( \frac{1}{d(v,w)\alpha} \), where \( d \) is the distance formula and \( \alpha \) is some coefficient.
Cost: We model the cost for entering a successor state in the search problem using a distance metric representing genetic expression variation between the current state and the successor state. Intuitively, the greater variation in the gene expression profiles of a cell A and a cell B, the higher cost this model assigns and then normalize it over all values. We proceed to incorporate this into our equation as a weight spread, and then normalize it over all values. We proceed to incorporate this into our equation as a weight spread.

Distance Metric
We use a high-dimensional Euclidean distance for costs in the search problem. The data used by Moignard et al provides gene expression levels of 46 genes for each of the 3935 single cells (figure 3). By computing the Euclidean distance across 46 dimensions (genes) for a pair of cells based on normalized expression values, we obtain a primitive metric for cell similarity that we use to compute transition cost (figure 4A). One downside to this approach is that it homogeneously weighs each of the differences across 46 genes. For instance, if gene \( j \) were crucial to cell function and controls differentiation from cell type PS to NP, a change \( \delta \) of this individual gene expression would be weighted equally with the same change \( \delta \) of an arbitrary gene in our model that has little to no effect on this transition. Therefore, we weight distance contributions of individual genes through the following method using a "Weight Spread."

Weight Spread
To prevent weighting all of our genes equally in our system, we weight our distances using the standard deviation of expression levels for a gene in each cell in our data set (figure 4B). The intuition for this is as follows: if the expression profile of a gene X has very little variance amongst all single cells in our data set, whenever we do have a change in expression for gene X during a transition, we can assume that it has a large effect on cell fate determination. To compute this weight spread value, we compute the standard deviation of our gene expression levels per gene, and then normalize it over all values. We proceed to incorporate this into our equation as a weight spread, and then normalize it over all values. We proceed to incorporate this into our equation as a weight spread.

\[
\text{Cost}(\text{cell}_1), \text{cell}_2) = \sqrt{\frac{(g_{11} - g_{12})^2}{\alpha_1} + \frac{(g_{21} - g_{22})^2}{\alpha_2} + \ldots + \frac{(g_{K1} - g_{K2})^2}{\alpha_K}}
\]

\[
\alpha = \frac{(\sigma_g)^2}{\sum_{i=1}^{K}(\sigma_j)^2}
\]

End State: We use an arbitrary termination state (END) that an erythroid or endothelial fate state (4G, 4GF) will deterministically enter. This allows each progenitor to find the shortest distance to a given final state once entering our known terminal states.

2.2 Random Walk Model
The next paradigm we use to model cell differentiation is a random walk model. This paradigm is fairly straightforward. For each progenitor cell, use a sampling function defined using either the Erdos Renyi or Kleinberg model to randomly choose one cell in the successor state to transition into. Repeat until a cell has reached a terminal state.

Start State: Every possible progenitor cell (PS, NP, HF).

Successors/Actions:

Erdos Renyi Model - The Erdos Renyi model sampler function chooses a cell state in the successor type to return at random with equal probability.

Kleinberg Model - The Kleinberg model sampler function differs from the Erdos Renyi sampler function in that it is not entirely random. The Kleinberg model sampler returns a cell state in the successor type with probability proportional to \( \frac{1}{d[v,w]^2} \) where \( d \) is the distance formula and \( \alpha \) is some coefficient.

We are optimistic that the Kleinberg sampler may return better results than the Erdos Renyi sampler, as cells are motivated to exert the minimum amount of energy they can in each transition. In this case we define energy in terms of genetic transitions and define our distance function as the Euclidean distance function described above, where each of the 46 gene expression levels in each cell state is an input to the Euclidean distance function.

End State
Figure 4: **Biplots of Data with Baseline UCS Implementation:**
(A) Biplot of original dataset provided by Moignard et al. Centroids of each cell stage visualized by clusters shown. (B) Biplot of a baseline UCS implementation for progenitor cell assignment. Shown are all primitive cells assigned into either erythroid (4G) or endothelial (4GF) by shortest paths to cells of either terminal type. Note the granularity of the “decision boundary” in the bottom right quadrant of the graph and that almost no primitive cells in the top left quadrant are assigned erythroid fates, which would require “leaving” the manifold and traversing the blank space in the middle of the biplot.

A cell reaches an end state in our random walk when it reaches a fully differentiated erythroid or endothelial cell fate (4G or 4GF).

3 RESULTS

To visualize the states and cell fate given 46 dimensions (genes), we use principle component analysis to project the data into the two most important dimensions of variability. We show in figure 7A the original HSC data categorized into stages of cell differentiation done by Moignard et al. In figure 7B, we show a baseline implementation of UCS that assigns all progenitor cells an erythroid or endothelial fate. We assign each progenitor cell (of types PS, NP, and HF) to an endothelial or erythroid fate based on the shortest path from it to any known endothelial (4GF) or known erythroid (4G) cell.

3.1 UCS - Erdos-Renyi Results

We sample 100 random graphs using a uniform, Erdos-Renyi style distribution. For each such random graph, we compute the shortest path from each progenitor cell to a given cell fate using UCS. We assign each progenitor cell to the terminal fate to which it was assigned in a majority of the samples (figure 7).

3.2 UCS - Kleinberg Results

We sample 100 random graphs using a distance-based, Kleinberg style distribution. We assign each progenitor cell to the terminal fate to which it was assigned (by UCS) in a majority of the samples (figure 8).
3.3 Random Walks - Erdos-Renyi Results

We sample 100 random walks from each progenitor cell using a uniform, Erdos-Renyi style distribution. Each random walk evolves a progenitor cell according to the time-based classification (figure 1) until it ends in either an erythroid (4G) or an endothelial (4GF) state. We assign each progenitor cell to the terminal fate to which it was assigned in a majority of the walks (figure 9).

3.4 Random Walks - Kleinberg Results

We sample 100 random walks from each progenitor cell using a distance-based, Kleinberg style distribution with a coefficient \( \alpha = 1 \). (B) Using a Kleinberg style distribution with a coefficient \( \alpha = 2 \).

Figure 7: Random Walks - Erdos-Renyi Implementation: The above biplot shows cell fate assignments based on the random walks begun at each progenitor that end in either an erythroid (4G) or an endothelial (4GF) state and therefore fate.

Figure 8: Random Walks - Kleinberg Implementation: The above biplot shows cell fate assignments based on the random walks begun at each progenitor that end in either an erythroid (4G) or an endothelial (4GF) state and therefore fate.

Figure 9: Random Walks - Kleinberg Implementation: (A) The above biplot shows cell fate assignments based on the random walks begun at each progenitor that end in either an erythroid (4G) or an endothelial (4GF) state and therefore fate using a Kleinberg style distribution with a coefficient \( \alpha = 1 \). (B) Using a Kleinberg style distribution with a coefficient \( \alpha = 2 \).
tion with a coefficient $\alpha = 1$ and 100 random walks with a coefficient $\alpha = 2$. We assign each progenitor cell to the terminal fate to which it was assigned in a majority of the walks for $\alpha = 1$ and for $\alpha = 2$ (figure 11).

4 DISCUSSION AND ANALYSIS

4.1 Label Analysis

To assess whether the predicted labels computed by UCS and through random walks under either the Erdos-Renyi or Kleinberg style models are proportionate, we compare the proportion of progenitor cells predicted to be erythroid (4G) to the empirical proportion of known erythroid (4G) cells out all terminal (erythroid and endothelial) cells in the dataset. We find an empirical proportion of $p_b = 0.56$. We find a predicted proportion of 4G labels for PS cells in the Baseline UCS implementation of $p_b = 0.16$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PS</td>
<td>0.26</td>
<td>0.26</td>
<td>0.56</td>
<td>0.48</td>
</tr>
<tr>
<td>NP</td>
<td>0.13</td>
<td>0.12</td>
<td>0.56</td>
<td>0.48</td>
</tr>
<tr>
<td>HF</td>
<td>0.17</td>
<td>0.17</td>
<td>0.56</td>
<td>0.48</td>
</tr>
</tbody>
</table>

As expected, the random walk paradigm using the Erdos-Renyi style distribution resulted in a predicted proportion of 4G labels ($P_{RW-ER} = 0.56$) approximately the same as the empirical proportion; this is due to the fact that once the random walk is at an HF cell, approximately 56 percent of the time, a uniform sampling of all erythroid (4G) and endothelial (4GF) cells will return an erythroid (4G) fate.

The random walk paradigm using the Kleinberg style distribution ($\alpha = 1$) resulted in a decline in the predicted proportion of 4G labels for each primitive type of PS, NP, and HF of 8 percentage points, and the random walk paradigm using the Kleinberg style distribution ($\alpha = 2$) resulted in a further decline of an additional 8 percentage points. This decline in the reachability of erythroid (4G) fates when sampling successors with probabilities inversely proportional to their distance away is commensurate with the high degree of Euclidean separation displayed by the primary known erythroid cluster from the rest of the developmental manifold (figure 4).

Surprisingly, shortest paths computed by UCS under either the Erdos-Renyi or Kleinberg style distributions resulted in virtually the same predicted proportion of 4G labels for each primitive type of PS, NP, and HF. In other words, the network model and distribution according to which edges were drawn for producing each random graph - whether uniform in an Erdos-Renyi manner or inversely proportional to distance in a Kleinberg manner - did not ultimately affect the terminal fates predicted for each progenitor cell. This reflects the possibility that the initial genetic distribution of a stem cell is the primary determinant of that cell’s terminal fate, and thus stochasticity in the initial distribution ultimately determines which embryonic stem cells differentiate into which specialized cells.

Overall, the predicted labels suggest that modeling cellular differentiation as a network-based search problem solvable through UCS produces results that are incompatible with empirically-observed proportions, as opposed to modeling it as a network-based random walk. That being said, whereas the proportion of labels informs the quality of a particular paradigm (search problem or random walk) and network model (Erdos-Renyi or Kleinberg) in producing equitable terminal outcomes in the aggregate, the quality of the labels - driven by the genetic transitions needed to achieve them - informs the quality of a particular paradigm and network model in reproducing known gene regulatory networks.

4.2 Average Transition Analysis

To compare the shortest paths computed by UCS and the random paths computed by random walks to known regulatory networks and transcription dynamics, we identify the average transitions in gene expression levels for all cell states. We separate results into two categories: PS, NP, and HF progenitor cells assigned to erythroid (4G) fates, and PS, NP, and HF progenitor cells assigned to endothelial (4GF) fates. For each category, we compute the average change in gene expression profiles between adjacent states (ex. PS → NP, NP → HF, etc.) over all generated random graphs and walks and report these transitions in heat maps. Here, we report heat maps for the Baseline UCS (figure 10), UCS - ER (figure 11), UCS - Kleinberg (figure 12), Random Walks - ER (figure 13), Random Walks - Kleinberg ($\alpha = 1$; figure 14), and Random Walks - Kleinberg ($\alpha = 2$; figure 15) implementations.

We track genes of interest in our model by comparing how UCS chose shortest paths and how random walks generated random paths under either distribution to known regulatory networks elucidated by Moignard et al. and earlier authors. In particular, Cdh1, Cdh5, Etv2, Gata1, Gfi1b, HbbbhH1, Ikaros, Myb, Nfe2, Runx1, Sfp1, and Tal1 are genes that have been shown to indicate erythroid (4G) fate determination, and Cdh5, Erg, Fli1, Hoxb4, Iga2b, Pecam1, Sox7, Sox17 are genes that have been shown to indicate endothelial (4GF) fate determination. The UCS baseline implementation successfully produces lower average transition values for Gata1, Gfi1b, HbbbhH1, Ikaros, Myb, Nfe2 in the NP to HF transition for cells that were eventually assigned to endothelial fates. For these same genes, we see higher average transition values for cells that were assigned to erythroid fates, confirming that modeling cellular differentiation as a search problem can successfully
reproduce much of the results of statistical analyses and biological assays into elucidating gene regulatory networks (e.g. those in Moignard et al.). Similarly, we see that Cdh5, Erg, and HoxB4 have a higher transition value from NP to HF, and Sox17 for HF to 4GF for cells that were eventually assigned to endothelial fates. For cells eventually assigned to erythroid fates, we see that Sox7, HoxB4, and Cdh5 all have strong negative transition values at some point along the process, indicating they likely are being suppressed for the formation of erythroblasts.

The UCS - Erdos-Renyi and UCS - Kleinberg implementations produced virtually the same solid results. Average transitions for those progenitor cells predicted to have endothelial (4GF) fates included striking increases in the cadherins (Cdh5), Hox factors (HoxB4), Sox factors (Sox7, Sox17), Pecam1, and Itga2b - genes all related to endothelial determination. However, the average transitions from HF included relative increases in the hemoglobin HbhbH1 and Nfe2, another erythroid determinant. Moreover, average transitions for those progenitor cells predicted to have erythroid (4G) fates were missing increases in known erythroid determinants HbhbH1, Gata1, Gfi1b, and Ikaros. Therefore, while both random UCS implementations were significantly faster than the baseline UCS implementation, they produced noisy results.

None of the UCS implementations successfully identify Erg as a primary driver of endothelial fate determination. By inspection of the biplots in figures 4A and 4B, we note that the progenitor PS, NP, and HF cells that are assigned an erythroid fate by UCS but also display increases in Erg transcription in the computed average transitions most likely do not enter into terminal erythroid states in the isolated cluster in the bottom left portion of the manifold (direction of decreasing Erg transcription), but rather into terminal erythroid states in the bottom right portion of the manifold (direction of increasing Erg transcription) obscured by other cell types.

Finally, each of the Random Walks implementations produced results that mostly reflected the greatest directions of genetic variance within the original dataset. For instance, in each of the three implementations and for each of the two predicted terminal fates, average transitions from HF always include a great increase in the hemoglobin HbhbH1; a quick reflection of the biplot at Figure 4A showing that HbhbH1 is a linear combination of the first two principal components, suggesting its importance. In fact, known erythroid (4G) cells have expression levels of HbhbH1 that are so drastically higher than those of the progenitor cells and known endothelial (4GF) cells that due to randomness, out of large number of random walk samples, many transitions incorporated in the average will reflect the drastic transition necessary to randomly walk to a known erythroid (4G) cell. In other words, the Random Walks implementations reflect the greatest directions of variance in the dataset in the aggregate, as opposed to the greatest directions of variance explored in locally searching for a particular terminal fate.

5 CONCLUSION

In conclusion, we demonstrate two possible network-based models for modeling cell fate differentiation. The search-based method with Uniform Cost Search produces average transition metrics that compare closely to known regulatory networks of transcription factors described by Moignard et al. Though finding the shortest paths for each progenitor cell through UCS is computationally intensive, the runtime of the algorithm can be dramatically reduced by replacing a complete successor graph with one that is randomly sampled according to some distribution (e.g. a uniform Erdos-Renyi style distribution or distance-based Kleinberg style distribution) at some expense of information quality. Since conducting biological assays for researching candidate gene regulatory networks is expensive, it may be preferable to spend more computational time finding more accurate results that could save larger costs in the future.

On the other hand, the Random Walks implementation is problematic to adapt for modeling the dynamics of cellular differentiation. Random sampling
Figure 11: **Analysis of UCS - ER Implementation:** (A) This heat map for our cells that have been assigned to endothelial fates (4GF) outlines the average transition change between cell stages per gene of interest. (B) Heap map for cells assigned to erythroid fates (4G).

Figure 12: **Analysis of UCS - Kleinberg Implementation:** (A) This heat map for our cells that have been assigned to endothelial fates (4GF) outlines the average transition change between cell stages per gene of interest. (B) Heap map for cells assigned to erythroid fates (4G).

Figure 13: **Analysis of Random Walks - ER Implementation:** (A) This heat map for our cells that have been assigned to endothelial fates (4GF) outlines the average transition change between cell stages per gene of interest. (B) Heap map for cells assigned to erythroid fates (4G).

Figure 14: **Analysis of Random Walks - Kleinberg Implementation** (**α** = 1): (A) This heat map for our cells that have been assigned to endothelial fates (4GF) outlines the average transition change between cell stages per gene of interest. (B) Heap map for cells assigned to erythroid fates (4G).
of successors - even sampling based on distance (even penalizing such distances drastically by tuning the Kleinberg distribution hyperparameter $\alpha$) - produces average transitions that ignore necessary local structure in the embryonic network.

We suggest using a search-based approach in revealing candidate gene regulatory networks using single-cell gene expression datasets. After all, modeling cellular differentiation as a network search problem manifests average expression level changes for genes of known interest that reflect experimentally determined transcription factors crucial for determination and thus may indicate new target genes for testing.

6 REFERENCES


7 GITHUB

https://github.com/sashaperigo/cs224w-final