**NIH Proposal Outline**

Six Page Limit – For Activity Codes R03, R13, R21, R36, SC2, SC3

1. **Introduction to Application**

The central goal of this proposal is to advance the state of the art in single-cell derived gene Boolean implication network construction and visualization.

1. **Specific Aims**
2. Development of a novel methodology to infer directed gene regulatory networks from single-cell RNASeq data.
   1. Evaluate the utility of different correlation and dependency metrics in the construction of directed regulatory networks. Possible metrics to consider: Pearson correlation, Spearman correlation, dCorr, MIC, and DREMI.
   2. Apply our novel inference network workflow to various subsets of single-cells of interest according to a pseudo-temporal tree of cell states to better understand the dynamic interactions of genes through a biological process of states and/or differentiation of cell types.
   3. Expand our analysis to activator-repressor-target gene relationships.
3. Visualize for hypothesis generation tool.
4. **Research Strategy**
   1. **Significance:** Gene dependence and correlation analyses have long been used to investigate the biological processes underpinning samples of interest. Recent work has been done regarding the susceptibility of traditional transcriptomic technologies to Simpson’s Paradox—the confounding of a mixture of signals that suggests a trend With the rise of single-cell RNAseq technologies, transcriptomics can now play a role in answering questions regarding tissue heterogeneity. This methodology promises a glimpse of the ‘rules of the game’ for a given progression across a dynamic cellular biological process. Pervious work from our lab performed similar inference of Boolean implication networks with microarray data (Yates). However, this methodology suffered, in part, from poor interpretability given the unknown cellular composition of the input datasets. A single-celled perspective in theory does not suffer from the averaging effects of bulk sample transcriptomics data and therefore derived implications in this work will not suffer from the same interpretation issues. Sub-aim 1.b, provides an exciting first look into the dynamics of Boolean implication of a given dataset and biological context. This work can further be used as a starting point for interesting future work, such as the generalization of implication detection in single-celled data from bivariate relationships to, the potentially more-interesting, gene triplet relationships.
   2. **Innovation:** To our knowledge no work has attempted to infer Boolean implication networks from single-cell RNASeq data. Furthermore, no other work has attempted to combine the works of Sahoo’s Boolean implication inference methodologies with Pe’er’s DREMI dependency metric. We believe that the combination of these two techniques will be able to detect potential bivariate gene implications that would otherwise be masked do to the rarity of certain cellular states in a given dataset.
   3. **Approach:** The specific aims of this proposal are presented in the order of intended execution. We believe that the completion of preceding aims will inform the completion of a given aim. In this section we will detail our intended approach to However, it should be noted that these aims still exhibit a separation of concerns, meaning that the success of one aim does not depend on the success of any other specific aim.

**Aim 1**

The first goal is to develop a novel methodology to infer regulatory gene networks from single-cell RNASeq data. To take full advantage of the resolution of this technology, the dataset(s) should include samples of likely different cellular subtypes. Therefore, tissue whose cells are suspected of going through a dynamic biological process such as proliferation, differentiation, or transition into malignancy may be of interest. Such a dataset would also be well suited for evaluation of aim 2. Furthermore, the number of samples per subtype will impact the accuracy of estimated gene dependency metrics. Taking cues from Marinov et al. who looked at various sized pools of single-cells subjected to the SMART-seq protocol and found that populations of 30-100 approach the information content of bulk tissue samples For this we will require a dataset…

Selecting the correct correlation and/or dependency metric(s) is an important sub aim to this work (Aim 1.a). Not only will different metrics limit the range of gene relationships that can be captured by our regulatory network, but each metric used will alter the interpretation of our resulting network. For these reasons we feel that it is necessary to encompass a broad range of metrics in our procedure, and comparing the regulatory networks that they each produce. We will also take each metric’s computational efficiency into consideration.

The DREMI metric [Eq. 4] is of particular interest to this investigation because it has been shown to expose functional relationships between variables whose joint probability is dominated by a seemingly independent signature (Dana Pe’er). To grasp the metric fully, consider the following definitions:

|  |  |
| --- | --- |
| Let and where |  |
| Let and |  |
| Let and |  |

Namely,andare vectors of random variables whileandare vectors of a sample’s X and Y components from the joint distribution .

Imagine that there may exist cell populations that are dominated by a particular cellular state or subtype (we’ll just say ‘state’ here for simplicity). Values in and may be dominated by a specific sub-range of the spans of and. If values in the dominant cell state are centralized around some (x, y) value, then many metrics may conclude that no relationship exists between genes X and Y. However, rare signatures for X and Y may offer discerning information regarding the functional dependence of genes X and Y. DREMI addresses this issue by subsampling according to an estimation of rather than . We’ll refer to this subsampling as along with the following definitions:

|  |  |
| --- | --- |
| Let and where |  |
| Let and |  |
| Let and |  |

DREMI de-convolution is achieved by estimating via a non-parametric diffusion kernel [Eq. 1] applied over bins of X values. Data is resampled according to this estimated conditional probability. And finally, mutual information [Eq. 2] is calculated for the down-sampled data. They show that this is equivalent to mutual information [Eq. 3] where every sample is weighted by [Eq. 4].

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |
|  |  | () |
|  |  | () |

The DREMI metric gives a reasonable measure for the strength of functional relationships between gene X and gene Y. The next step is to extract inference wherever possible. In other words, we want to turn our undirected network into a directed one. A number of previous works attempt to infer directionality in transcriptome data (Sahoo 2008). Previous work from our lab have successfully utilized the approach used by Sahoo et al. for the ‘fuzzification’ and directionality inference of relationships extracted from microarray data (Yates). We intend to use the same basic methodology applied to the down-sampled technique described in (Pe’er et al).

For a scatterplot of , we start by discretizing the subsampled data as described in (Yates) by labeling a given gene expression value as either ‘high’ or ‘low’. This is done by first ordering all observed values of from smallest to largest and fitting a step function to those ordered values using StepMiner—an algorithm presented in (Sahoo, Extracting binary signals from microarray time-course data). These step functions aim to minimize the mean squared error (MSE) using an adaptive regression process. Next, the average of the high and low steps in a fitted step function serve to be the decision threshold between fuzzy ‘high’ and ‘low’ labels for gene X.

To discretize , we may be able to take advantage of the P(Y|X) normalization effects that occur in the subsampling proposed by (Pe’er). In our lab’s original work, step function fitting was chosen, in part, because the minimization of MSE can still find a suitable boundary decision line in the presence of outliers. However, Pe’er’s conditional probability normalization resampling methodology can serve to remove outliers. We propose that a simpler boundary decision can be found:

|  |  |  |
| --- | --- | --- |
|  |  | () |

After discretization, we continue Sahoo’s workflow to test for sparsity amongst the quadrants formed by the decision boundaries for X and Y. This is done by testing sample counts in a quadrant against the null hypothesis of uniformity in the distribution of samples. For instance, consider the test for the quadrant corresponding to low values of both X and Y. Letting , , , and be the number of samples categorized with X and Y labels ‘low’ and ‘low’, ‘low’ and ‘high, ‘high and ‘low’, and ‘high and ‘high respectively. Sparsity of the low-low quadrant is determined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | () |

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

|  |  |
| --- | --- |
| , for thresholds and | () |

Analogous calculations are performed to determine , , and . Now let, , , , and be Boolean variables that are true when, for a particular biological context, gene X or Y is highly expressed or not highly expresses (note that and ). Depending on the sparsity profile of the quadrants we propose that the following implications can be drawn:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | Figure | Conclusion |
| 1 | 1 | 1 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1111.png | No implication |
| 1 | 0 | 1 | 0 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1010.png | No implication |
| 0 | 1 | 0 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:0101.png | No implication |
| 1 | 0 | 0 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1001.png |  |
| 0 | 1 | 1 | 0 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:0110.png |  |
| 1 | 1 | 0 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1101.png |  |
| 1 | 0 | 1 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1011.png |  |
| 0 | 1 | 1 | 1 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:0111.png |  |
| 1 | 1 | 1 | 0 | Macintosh HD:Users:Brian:Documents:Research:CandidacyProposal:Figures:1110.png |  |

The set of resulting, gene pair-wise Boolean implications constitute a Boolean implication network between ‘fuzzified’ expression values in a given biological context.

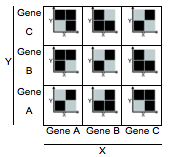


Figure - Example implications (Left), and implication network (Right) for a set of three genes A, B, and C.

Note that the conditional probability subsampling methodology proposed does not guarantee symmetric implications about X and Y. Non-symmetric patterns may be of interest biologically—potentially indicating the presence of a rare expression pattern.

One shortcoming that may need to be addressed is our approach’s oversimplification of general biological systems. Gene interactions are rarely independent of one another. It is a strong assumption to claim that functional relationships will be identifiable when considering pairs of features at a time. For this reason we generalize our approach in sub-aims 1.b and 1.c to account for different types of conditional dependence between gene X and gene Y.

**Aim 1.b- Pseudotemporal Model**

Next, we attempt to decouple functional relationships between gene X and gene Y from conditional dependence upon a factor that Trapnell et al. refer to as ‘pseudotime’—“a quantitative measure of progress through a biological process” (Trapnell Monocle). The interplay of genes is known to change as a cell progresses through dynamic biological processes such as differentiation, cell cycle, or oncogenic transformation. Our approach may not be able to identify strong relationships between X and Y when considering cells from multiple cellular states if the true, underlying relationships between X and Y in those states differ. To overcome this convolution of signals through pseudotime, we combine sample clustering with the temporal ordering output of Monocle. We plan to cluster samples in 2-dimensional independent component analysis (ICA) space using hierarchical clustering.

Monocle orders single-cells along an inferred smooth transition function in Euclidean gene expression space. Here represent the true set of expression values between states, and captures biological and technical noise. Monocle estimates with the following methodology, which was first introduced by Magwene et al. First, independent component analysis (ICA) is performed on the data, and fitted with a minimum spanning tree (MST) in 2-dimensional independent component space. The diameter of the MST is taken as an estimate of . The algorithm goes on to find potential orderings of samples in the data relative to . A PQ tree is created with a single Q node denoted QMain. A PQ tree defines a family of orderings of descrete elements in a set; it is a tree with two types of nodes—a Q node whose children are ordered (although reversible), and a P type node where children are permutable. All vertices along the diameter of the MST with degree greater than 2 are deemed ‘indecisive’, and ‘decisive’ otherwise. Find the ‘indecisive backbone’ of the diameter—the longest sequence vertices for which the endpoints are indecisive. Add any decisive vertex along this indecisive backbone to QMain in an ordered fashion. Then for any indecisive vertex on the indecisive backbone add a P node to the tree and add the indecisive vertex as a child of that P node. Recursively apply the entire algorithm to each branch of the indecisive vertex adding. Possible orderings of samples in pseudotime are given by those orders extractable from the PQ tree.

Using the ordering proposed by the PQ tree with the shortest total distance in component space to define our samples’ ordering in pseudotime, we can define an order between sample clusters. We’ll let be an ordered set of the indices output by Monocle that orders our samples in pseudotime. Cluster can then be ordered by We will say that a given clustering agrees with the MST output by Monocle if the variance of per cluster is sufficiently small. We purposely supply this crude definition of ‘agree’ because the clustering step will be a data-driven process and highly customizable by method and parameterization. We do believe that the clustering achieved by hierarchical clustering and the ordering of individual samples by Monocle will largely agree because if two samples are assigned to the same cluster they are relatively close to each other and Monocle will therefore likely assign indices that are relatively close to each other as well.

At this point we will assume that our clusters constitute different cellular states separated by different progressions through pseudotime. By constructing Boolean implication networks by the approach outlined for aim 1 for each individual cluster, we are in a sense looking at relational dependencies between genes at different slices of pseudotime. More interestingly, maybe, are the Boolean implication networks that result from considering samples that belong to pairs of adjacent clusters according to and Monocle’s MST. Effectively we can construct the rules of genetic interplay for both major cellular states in a dynamic biological process as well as for the transitions between those states.

**Aim 2 –Visualization**

We will build an interactive visualization that will allow investigators to navigate the networks we construct in aim 1—complete with appropriate organizational and informational tools that can help the user with gene regulation oriented hypothesis generation.

First we wish to develop a concise visual encoding capable of conveying all classes of Boolean implication between gene pairs possible as output of aim 1 of this proposal. A traditional Boolean implication network visualization, as seen in Fig. 1, requires nodes (‘high’ expression and ‘low’ expression labels per gene). Such a separation can complicate hypothesis generation.

Our application will be provided as an R package developed using the web interface framework Shiny. Scatterplots and heatmaps will be realized using the ggplots package and network data structures and visualizations will be handled using the igraphs package.

Visual encodings:

* Edge thickness will be proportional to DREMI .
* Pairs of directed Boolean implication classes will be encoded according to the following chart:
* Node color will denote user-specified gene groupings.

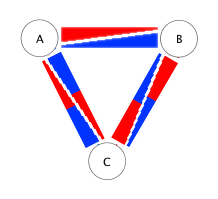
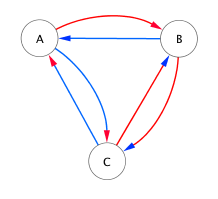


Figure 2- Concise visualization of example implication network introduced in figure 1 (Left). Colors denote high and low labels of genes in an implication relationship. Colors of arrow stems map to the upstream gene in an implication while arrowheads map to the downstream gene in an implication. And our proposed visual encoding of edges (Right) which captures the same information.

Use Cases:

* Use biomaRt to query selections against public genetics databases (Kegg, David)
* Filter by edge weight, edge class, or node weight.
* Will allow for custom grouping of samples and/or genes.
* Will allow for custom sorting of samples and/or genes.
* Visualize data scatterplots (either original or DREMI resampled) of selected implications.
* Heatmap

**Datasets**

We will focus on stem cell differentiation datasets with a large number of samples to test our analytics. GSE60749 is an examination of pluripotent stem cells (PSCs) in mus musculus brain tissues. Kumar et al. find that there were two clusters 1 composed of 98% of the samples and the other only 8% or 14 individual cells. This breakdown may constitute a rare cellular state on which we can examine

GSE64016 provides 460 human embryonic stem cells (hESC)—213 H1 single cells and 247 H1-Fucci labeled single cells.

GSE65525 provides ~3000 UMI-barcoded, differentiating, embryonic mouse stem cells from a massively parallelized microfluidics-based single-cell sequencing technology called DropSeq.

**Future Work – Gene Triplet Implications**

Sub-aim 1.c attempts to do a similar decoupling. But instead of decoupling the functional relationships between gene X and gene Y from conditional dependence upon pseudotime, we consider the conditional dependence upon a third gene Z. This is an attempt to model the activator-target-repressor (X-Y-Z respectively) relationship commonly found in biological pathways. A natural Woolf and Wang present an exhaustive approach to check all gene triplet signatures against a model finding such conditional dependencies.

Then we can start talking about the persistence