DIISCO: Dynamic Intercellular Interactions in Single Cell transcriptOmics

Cameron Park[*1] **Shouvik Mani**[*2] **Satyen Gohil**[3] **Katie Maurer**[4] **Catherine J Wu**[3 4] **Elham Azizi**[1]

Abstract

Inferring cellular interactions is essential in understanding the coordination of biological processes in normal physiology and disease. Current methods predict cell-cell interactions using single-cell RNA-sequencing (scRNA-seq) data based on expression of known receptor-ligand pairs. There are two main limitations to this approach. First, this requires a comprehensive database of known protein-protein interactions. Any unknown interactions (especially with rare cell types) will not be well-captured in the predictions. Second, these algorithms do not take into account the dynamic nature of cell-cell interactions which is crucial to consider in longitudinal single-cell data where cells are exposed to various conditions or treatments. Here, we present DIISCO, a method to characterize dynamic cellular interactions in longitudinal scRNA-seq data. Our method uses a gaussian process regression network to infer interactions between cell types according to changes in cell type proportions over time and incorporates prior knowledge on receptor-ligand interactions. We present preliminary results of this method applied to three datasets: CAR-T cells co-cultured with a leukemia cell line, simulated data, and T cell states from relapsed leukemia patients treated with adoptive cell therapy.

1. Introduction

Understanding the function of diverse cell types requires studying their dynamics and how they evolve with perturbations such as treatments. scRNA-seq is a powerful technology for elucidating diverse cell types and states. However, there is a strong demand for computational methods that are capable of integrating single cell data across multiple time-points, especially in longitudinal clinical data where sample timing is variable across patients. Importantly, to elucidate underlying mechanisms in complex systems such as the tumor microenvironment, we also need to learn the intricate interactions between cell types that orchestrate progression of disease or response to treatments. For example, in the context of immunotherapy, uncovering crosstalk between tumor and immune cells can help us understand immune dysfunction mechanisms and the mode of action of therapies.

Current methods for predicting cell-cell interactions in scRNA-seq data rely on existing databases of interacting protein complexes, and they predict interactions based on expression levels of known receptor-ligand pairs (Efremova, 2020). These protein complexes can have vastly different effects in different cell types, and this context-dependent nature is not considered by current methods. Reliance on existing databases also means that novel interactions and rare cell types may not be accurately predicted by the method due to the lack of existing studies characterizing these systems. Furthermore, single cell interaction methods do not consider temporal dependencies, which are essential in disease progression where patients are exposed to different treatments at various time points (Argelaguet, 2021). With longitudinal sequencing becoming more popular and readily available, there is a need for a method that can capture these dynamic changes. Complex systems such as the tumor microenvironment are ever-changing, especially as patients undergo therapy, with different cell types being recruited to the area and interactions between these cells leading to diverse treatment responses. Characterizing the dynamic nature of these interactions will help us understand the underlying mechanisms of current immunotherapy and guide new, more effective, therapies.

2. Methods

2.1. Problem Definition and GPRN Framework

Suppose we have data obtained from longitudinal scRNA-seq over a set of T non-uniformly spaced time points

[*]Equal contribution [1]Department of Biomedical Engineering and Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA [2]Department of Computer Science and Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA; Department of Medical Oncology, Dana- Farber Cancer Institute, Boston, MA 02215, USA [4]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA; Harvard Medical School, Boston, MA 02115, USA. Correspondence to: Cameron Park <cyp2111@columbia.edu>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

$t \in \mathbb{R}^{T}_{0+}$. Let $y(t_i) \in \mathbb{R}^C$ denote zero-centered proportions of C cell types at time t_i. The dataset $D_{T \times C} = \{y(t_i) : i = 1, ..., T\}$ records cell type proportions over time. Additionally, let $\Lambda \in \mathbb{R}^{C \times C}_{0+}$ denote a matrix of regularization penalties encoding prior knowledge on the strength of interactions between cell types – a low $\Lambda_{k,k'}$ signifies a strong expected interaction between cell types k and k'.

We model the cell type proportions $y(t_i)$ using a gaussian process regression network (GPRN) (Wilson et al., 2012).

$$y(t_i) = W(t_i)[f(t_i) + \sigma_f \epsilon] \tag{1}$$

Here, $W(t_i)$ is a $C \times C$ matrix of trainable parameters such that each entry over time is an independent gaussian process: $W_{k,k'}(t) \sim GP(\phi, K_W)$, i.e. $(W_{k,k'}(1), ..., W_{k,k'}(T)) \sim N(0, K_W)$. $f(t_i)$ is a $C \times 1$ vector of latent functions representing cell type proportions at time t_i, and is initialized by fitting independent gaussian processes (GP) for each cell type in y, so that $f_k(1), ..., f_k(T)$ approximate cell type k's proportions over time $y_k(1), ..., y_k(T)$. K_W is a $T \times T$ kernel matrix with length-scale θ_W. z is i.i.d. $N(0, I)$ white noise, scaled by the noise hyperparameter σ_f. θ_f is the length-scale for the independent GPs used for the $f(t)$'s.

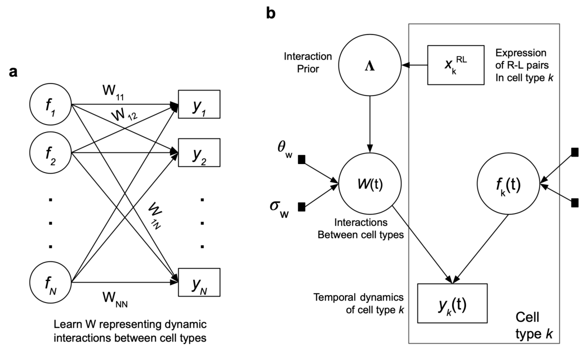

Figure 1. Left: General model design and structure. Right: Plate model denoting different variables used in model formulations. Square boxes represent observed variables, circles represent latent variables. θ and σ are hyperparameters used in defining the kernels for f and W.

By setting the $f(t)$'s to independent GPs fit to the proportions of each cell type, the model effectively predicts cell type proportions $y(t_i)$ from all other cell type proportions $f(t_i)$. Thus, the parameter $W_{k,k'}(t_i)$ can be interpreted as the contribution of cell type k''s proportion in predicting cell type k's proportion at time t_i. In other words, $W_{k,k'}(t_i)$ quantifies the strength of interaction between cell types k and k'.

Notably, with trainable parameters $W(t_i)$ and fixed latent functions $f(t_i)$ defined as above, a solution of $W(t_i) = I$ approximates cell type proportions $y(t_i)$ accurately, but provides no insight into the interaction dynamics between cell types. To avoid this trivial solution, we use regularization to penalize self-interactions and guide the model towards inferring cell type interactions known a priori, while learning from the data simultaneously. The regularization is equivalent to placing a prior on the model weights with higher variance for known interacting cell types:

$$W_{k,k'}(t) \sim GP(\phi, K_W) \tag{2}$$

$$\phi \sim N\Big(0, \frac{1}{\Lambda_{k,k'}}\Big) \tag{3}$$

In particular, the regularization penalty matrix $\mathbf{\Lambda}$ should have large entries along the diagonal to penalize self-interactions, as well as large entries for known non-interacting cell types. All other entries, especially those corresponding to known interacting cell types, should be set to low values.

We train the model to maximize the evidence lower bound (ELBO) using the scalable gaussian process regression network (SGPRN) framework (Li et al., 2020), with its loss modified to apply our custom regularization penalty:

$$\begin{aligned} L = KL(q(\boldsymbol{W})||p(\boldsymbol{W})) + KL(q(\boldsymbol{F})||p(\boldsymbol{F})) \\ - \mathbb{E}_q[\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{F})] + \|\Lambda \circ \boldsymbol{W}\|_2 \end{aligned} \tag{4}$$

where $\boldsymbol{W}, \boldsymbol{F}, \boldsymbol{Y}, \boldsymbol{X}$ are tensors of concatenated $W(t_i), f(t_i), y(t_i), t_i$, respectively.

The trained model predicts cell type proportions $\hat{y}(t_i)$ in the range $[t_1, t_T]$. Prediction intervals are generated for $\hat{y}(t_i)$ by sampling $W(t_i)$'s and $f(t_i)$'s from the trained GPRN, using equation (1) to predict $\hat{y}(t_i)$, and taking the top and bottom percentiles of the predictions. Most importantly, the learned $W(t_i)$'s can be visualized over time and interpreted to understand dynamic cell-cell interactions.

2.2. Prior on W

To better guide the model, we built a prior based on gene expression of known receptor-ligand (R-L) pairs. R-L pairs were taken from cellphoneDB (Efremova, 2020). After filtering out genes with low expression, gene expression was z-scored across all cells to account for differences in capture rate in sequencing technology. For each cell type, average z-scored expression μ_g for each gene g was calculated as well as standard deviation σ_g of expression for each gene. Within each celltype, genes with $\mu = 0$ and $\sigma = 0$ were disregarded as well as any genes with $\mu < 0$ or high variance. High variance for gene g is defined as $\sigma_g > mean(\boldsymbol{\sigma}) + 2std(\boldsymbol{\sigma})$ with $\boldsymbol{\sigma}$ being the matrix of standard deviations for each cell type-gene pair. To then build an interaction matrix, $\mathbf{\Lambda'}$, between cell types, each $\Lambda'_{k,k'}$ is defined as the $max(L_k + R_{k'})$, where L_k is ligand expression in cell type k, and $R_{k'}$ is receptor expression in cell type k'. To create the regularization penalty for W, we want highly-interacting clusters (which would have high scores in $\mathbf{\Lambda'}$) to have the lowest penalty in $\mathbf{\Lambda}$. Therefore, we defined $\mathbf{\Lambda} = max(\mathbf{\Lambda'}) - \mathbf{\Lambda'}$. We then set all diagonal values (where $k = k'$) to an order-of-magnitude higher than the highest penalty given in off-diagonal values.

2.3. Data Preparation and Model Setup

2.3.1. CAR-T and MEC-1

This dataset consists of 3 different experiments. CAR-T cells (with either the 41BB or CD3Z domain) were co-cultured with chronic lymphocytic leukemia cells (from MEC-1 cell line) over a 24-hour period and sequenced at 10 different non-uniform time points. The non-uniform time points are more realistic for comparison with real patient data. The main differences between experiments were the times at which they were sequenced and the domain used in engineering the CAR-T. Gene expression was used to annotate metaclusters into 4 groups: cancer cells, exhausted T cells, activated T cells, and other CD8+ T cells. We then calculated sample proportions in each of the metaclusters as the model input for each of the experiments. UMAP visualization of all cells with metacluster assignments as well as the proportions over time are shown in Fig. 2.

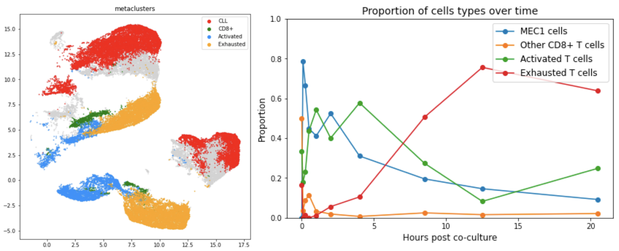

Figure 2. Left: UMAP with metacluster assignments based on gene expression. Right: Proportions over time for each metacluster.

2.3.2. Simulated data

To test the accuracy of this model in inferring complex interactions, we simulated proportions for five cell types over time with a 5×5 ground truth W matrix. The average and dynamic W_{true} are shown in Fig. 3. Specifically, out of 25 interactions terms, 21 were sampled from a $GP(0, K_W)$ and 4 were sampled from a $GP(\mu, K_W)$, where μ was set to various non-linear functions of the input to simulate 3 different types of dynamic interactions: interactions that disappear over time, interactions that grow stronger over time, and transient interactions that only persist for a short period of time. We believe these types of interactions will be important in the context of immunotherapy.

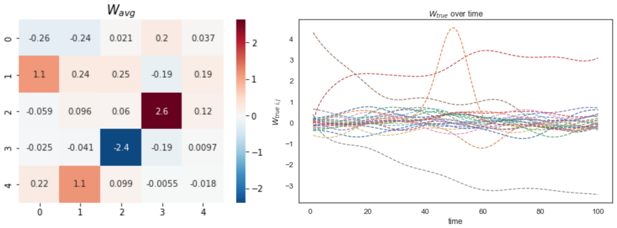

Figure 3. Left: Average of W_{true} matrix. Right: Dynamic W_{true} with four dominant interaction terms.

2.3.3. T cell states in CML

Bachireddy et al. used an independent gaussian process model on different T cell metaclusters in relapsed chronic myeloid leukemia (CML) patients who received donor lymphocyte infusion (DLI) treatment (Bachireddy P, 2021). They describe two distinct subpopulations of T cells with opposing dynamics in responders. MC3 displayed a terminally exhausted phenotype and contracted in responders after treatment. The second population, MC1 and MC2, displayed markers more representative of a progenitor exhausted phenotype and expanded in responders after treatment. We tested DIISCO on this dataset, limiting to responder cells and using the original metacluster assignments to increase statistical power and preserve batch correction efforts. The metacluster assignments can be seen in the UMAP shown in Fig. 4

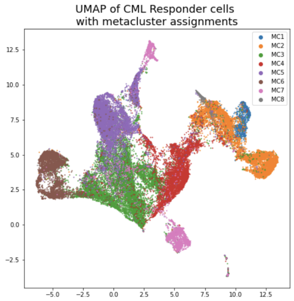

Figure 4. UMAP visualization colored by T-cell metacluster assignments for responder CML patients. Metacluster assignments based on Bachireddy et al.

3. Results

3.1. CAR-T and MEC-1

In the CAR-T system, there are only two major cell types present, and we expect a negative interaction between the MEC-1 and CAR-T cells. DIISCO is able to recover this interaction in the average W matrix, as shown by the strong negative weights between the MEC-1 and Exhausted T cell subsets in Fig. 5. This data may not be well-characterized by methods like cellphoneDB, because the CAR-T receptor is not captured in the sequencing data and therefore the interaction would not be in the reference database. However, our model can still predict interactions that align with the biological system.

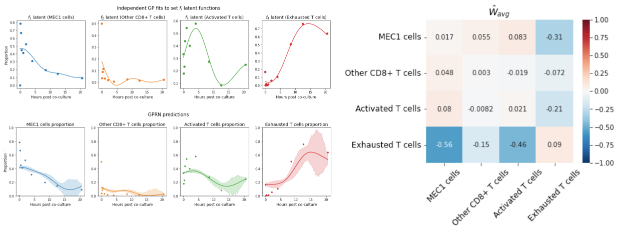

Figure 5. Left, top: Latent functions modeled by independent GPs. Left, bottom: Mean predictions and confidence intervals for each cell type proportion, with points representing actual proportions. Right: average W matrix of interactions learned by the model

3.2. Simulated data

We used simulated data to evaluate the benefit of regularization in the presence of more complex interactions. As shown in Fig. 6, DIISCO recovers the four dominant interactions present in the dynamic W_{true} matrix. When we compared a multi-level regularization penalty to regularizing only the diagonal terms in W, we observed that while both were able to capture the significant interactions between clusters fairly well, the multi-level regularization did a better job at limiting false positive interactions.

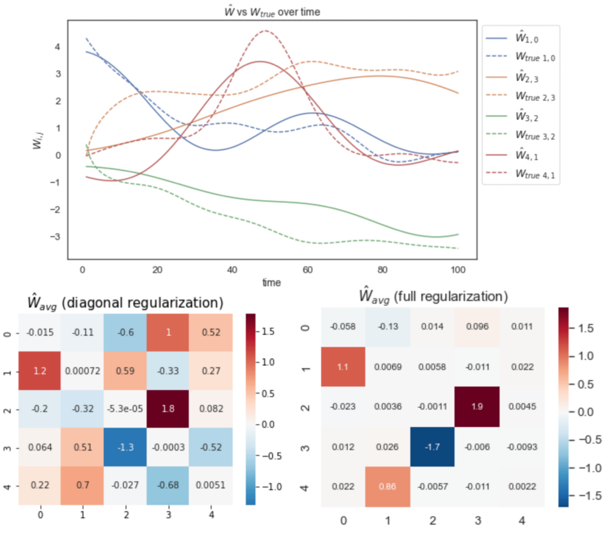

Figure 6. Top: Learned dynamic W compared with the true dynamic W for simulated interactions. Bottom, left: Average learned W using only diagonal regularization penalties. Bottom, right: Average learned W using regularization penalties on off-diagonal non-interacting cluster pairs.

3.3. T cell states in CML

Bachireddy et al. found 3 metaclusters (MC1, MC2 and MC3) with distinct dynamics in patients who responded to DLI treatment. MC1 and MC2 were shown to expand after treatment and MC3 was shown to contract. Fig. 7 shows the regularization penalties used for MC1-MC7 as well as model outputs for this system. Comparing our model to the independent gaussian process model used in the original study, we are able to recover similar learned proportions. Unlike the original paper, we are also able to analyze the weights of our model to understand how these different cell types interact over time. We see strong negative interactions between MC3 and MC1/MC2 and this interaction grows stronger after treatment also shown in Fig. 7.

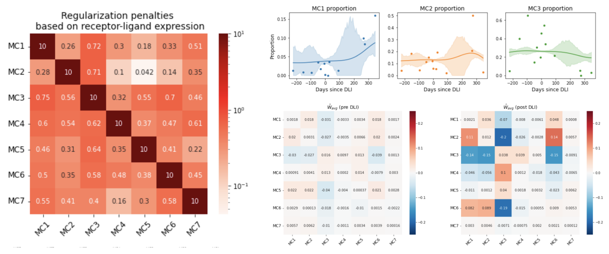

Figure 7. Left: Regularization penalties on W used to encode prior knowledge on receptor-ligand complexes. Right, top: Predicted proportions (line) compared to true proportions (points) for MC1, MC2, and MC3. Right, bottom: Average learned W before and after DLI treatment. Interactions between MC1/MC2 and MC3 grow stronger post-DLI.

4. Conclusion and Applications

As longitudinal scRNA-seq data becomes more readily available, our method can be applied and generalized to different cell types and diseases. We show how this method can be used in three different datasets of increasing complexity. We show the benefit of adding a regularization penalty that incorporates gene expression data, as well as the increased power in using well-characterized metaclusters. Future additions to the method include adding a prior to account for variable sample sizes. We are also expanding DIISCO to look at underlying gene patterns that may be guiding the learned dynamics, giving us a deeper understanding of the potential underlying mechanisms involved in the interactions. We are currently working on applying this method to a larger dataset of relapsed CML patients who undergo DLI. By expanding interactions to include all cell types present in the data, we hope to better understand the mechanisms underlying DLI and its success in over 70% of relapsed CML patients as a means to better understand relapsed AML (Greiner, 2020). Relapsed AML patients have fewer treatment options, and while DLI is standard-of-care for these patients, less than 20% of patients respond positively to treatment (Felicitas Thol, 2020). By comparing cell-cell interactions in these diseases, we hope to differentiate the shared and unique mechanisms that may be underlying their response or resistance to therapy.

References

Argelaguet, R., C. A. S. O. e. a. Computational principles and challenges in single-cell data integration. pp. 1202–1215. Nature Biotechnology, 10 2021. doi: 10.1038/s41587-021-00895-7.

Bachireddy P, Azizi E, e. a. Mapping the evolution of t cell states during response and resistance to adoptive cellular therapy. Cell Reports, 11 2021. doi: 10.1016/j.celrep.2021.109992.

Efremova, M., V.-T. M. T. S. e. a. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. pp. 1484–1506. Nature Protocols, 2020. doi: 10.1038/s41596-020-0292-x.

Felicitas Thol, A. G. Treatment of relapsed acute myeloid leukemia. pp. 2456–2462. Current Treatment Options in Oncology, 6 2020. doi: 10.24963/ijcai.2020/340.

Greiner, J., e. a. Immunological and clinical impact of manipulated and unmanipulated dli after allogeneic stem cell transplantation of aml patients. pp. 2456–2462. Current Treatment Options in Oncology, 6 2020. doi: 10.24963/ijcai.2020/340.

Li, S., Xing, W., Kirby, R. M., and Zhe, S. Scalable gaussian process regression networks. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2456–2462. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/340. Main track.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z. Gaussian process regression networks. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1139–1146, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.