Simultaneously aligning cells and features of single-cell multi-omic datasets with co-optimal transport

Anonymous Author(s)
Affiliation
Address
email

Abstract

Availability of different single-cell multi-omic datasets provide an opportunity to study various aspects of the genome at the single-cell resolution. Jointly studying multiple genomic features can help us understand gene regulatory mechanisms. Although there are experimental challenges to jointly profile multiple genomic features on the same single-cell, computational methods have been develop to align unpaired single-cell multi-omic datasets. Despite the success of these alingment methods, studying how genomic features interact in gene regulation requires the alignment of features, too. However, most single-cell multi-omic alignment tools can only align cells across different measurements. Here, we introduce SCOOTR, which aligns both cells and features of the single-cell multi-omic datasets. Our preliminary results show that SCOOTR provides quality alignments for datasets with sparse correspondences, and for datasets with more complex relationships, supervision on one level (e.g. cells) improves alignment performance on the other level (e.g. features).

1 Introduction

Recent experimental developments have enabled us to measure various aspects of the genome, such as gene expression, chromatin confirmation, chromatin accessibility and methylation, at the single-cell resolution [1–4]. Studying the multiple views of the genome together can allow biologists to learn how they interact to regulate cellular processes. Although we can experimentally combine some measurements on the same single-cell using co-assays, for most measurement combinations, there are no co-assays available [4]. Moreover, simultaneous profiling of multiple features can yield more noisy data than single-omic experiments [5]. As a result, various computational methods [6–12], including the ones based on optimal transport theory [9–11], have been developed to successfully align single-cell measurements from non-co-assay (i.e. unpaired) experiments.

Despite the success of these methods, studying cross-modality feature relationships also requires the alignment of features. Due to the number of features and the complexity of their relationship, we need new computational approaches to infer these alignments. Unfortunately, most single-cell alignment methods can only yield alignments on the cell level, with the exception of bindSC [12]. Although bindSC performs both cell and feature alignments, it requires prior knowledge of feature relationships using a gene activity matrix. This gene activity matrix is computed between gene expression features and the chromatin accessibility or methylation signals in the neighborhood of these genes. Therefore, it can only work with a few measurements (like chromatin accessibility and methylation) that have known relationships with gene expression and ignores most intergenic regions.

We introduce SCOOTR, which simultaneously aligns both the cells and the features of unpaired single-cell multi-omic datasets in a modality-agnostic manner and without systematically ignoring intergenic regions, using co-optimal transport. Our results demonstrate that SCOOTR can yield quality alignments for both cells and features between datasets with sparse correspondences. For datasets with more complex relationships, supervision on one level (e.g. cell-type alignments) improves alignment performance on the other level (e.g. features, or vice-versa).

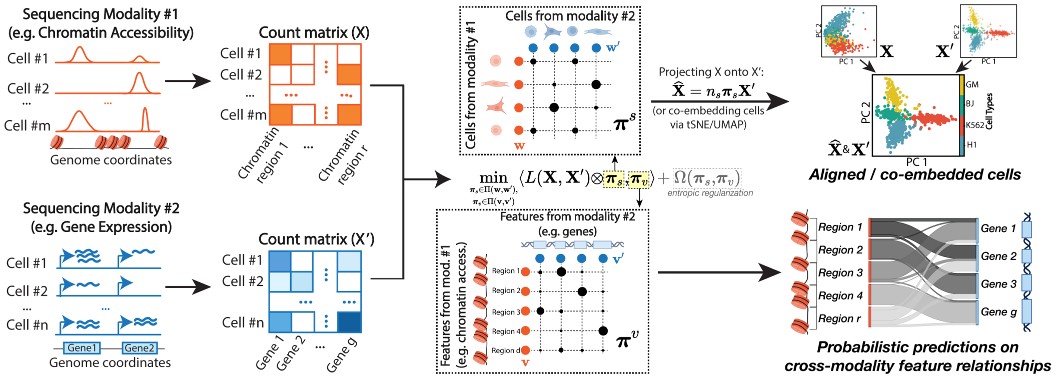


Figure 1: **Overview of the SCOOTR method on SNARE-seq dataset** Given two unpaired single-cell datasets with different genomic measurements (e.g. chromatin accessibility and gene expression), SCOOTR simultaneously solves for two probabilistic correspondence matrices: one between features, and one between cells across the two datasets.

2 Method

We first give a brief introduction to optimal transport and explain how the existing optimal transport-based single-cell multi-omic alignment methods discard features during alignment. We then introduce the SCOOTR framework.

2.1 Background on Optimal Transport

Optimal transport is a mathematical framework for relating probability distributions or discrete measures to one another. Here, we focus on discrete measures because we work with sequencing datasets that contain empirical measurements on a finite set of samples. Consider two datasets with n and n' data points in each, represented by matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ and $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}]^T \in \mathbb{R}^{n' \times d}$. We let $\mu = \sum_{i=1}^{n} w_i \delta_{\mathbf{x}_i}$ and $\mu' = \sum_{j=1}^{n'} w'_j \delta_{\mathbf{x}'_j}$ be two empirical distributions related to their samples. Here δ_{x_i} is the Dirac measure, the probabilities placed on data points are non-negative, $w_i \geq 0, w'_j \geq 0$, and sum up to one for each dataset, $\sum_{i=1}^{n} w_i = 1 = \sum_{j=1}^{n'} w'_j$. We refer in the following to $\mathbf{w} = [w_1, \ldots, w_n]^\top \in \Delta_n$ and $\mathbf{w}' = [w'_1, \ldots, w'_{n'}]^\top \in \Delta_{n'}$ as sample weights vectors that both lie in the simplex.

Given a cost function $\mathbf{L}(\mathbf{x}_i, \mathbf{x}'_j)$ that describes how "expensive" it is to match one data point ($\mathbf{x}_i$) with another ($\mathbf{x}'_j$) across the two datasets, Kantorovich formulation of optimal transport sets out to find an optimal coupling $\boldsymbol{\pi}$ that attains:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{w},\mathbf{w}')} \sum_{i,j} \mathbf{L}(\mathbf{x}_i, \mathbf{x}'_j) \boldsymbol{\pi}_{ij} \tag{1}$$

$$\text{such that } \Pi(\mathbf{w}, \mathbf{w}') = \{\boldsymbol{\pi} | \boldsymbol{\pi} \geq \mathbf{0}, \boldsymbol{\pi} \mathbf{1}_{n'} = \mathbf{w}, \boldsymbol{\pi}^\top \mathbf{1}_n = \mathbf{w}'\}. \tag{2}$$

Here, the coupling $\boldsymbol{\pi}$ holds the alignment probabilities between each pair of data points across the two datasets to optimally transform one into the other and Equation 2 defines the set of linear transport constraints. Most of the practical applications of optimal transport includes an entropic regularization over the coupling matrix to split the alignment probabilities across multiple samples and also to make the optimization computationally more efficient. For more detailed background on optimal transport, we refer readers to Villani, 2008 [13] (for theory), and Peyré and Cuturi (2019) [14] (for computational aspects).

2.2 Previous Optimal Transport-Based Single-cell Alignment Methods

Previously, three single-cell multi-omic alignment algorithms have been proposed based on optimal transport: SCOT [9], Pamona [10], and SCOTv2 [11]. In a single-cell multi-omic alignment task, the datasets to be aligned contain measurements from different modalities (with potentially different number of features d and d'). Performing alignment using the formulation in Equation 1 would require defining a cost function over samples of different metric spaces, which is not possible. To get around this, SCOT [9] used Gromov-Wasserstein optimal transport, which extends the formulation in 1 with a cost function defined over intra-domain sample distances, making it amenable to use for multi-omic datasets:

$$GW(\mathbf{w}, \mathbf{w}') = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{w}, \mathbf{w}')} \sum_{ik}^{n} \sum_{jl}^{n'} L(D^X_{ik}, D^{X'}_{jl}) \boldsymbol{\pi}_{ij} \boldsymbol{\pi}_{kl} \tag{3}$$

$$\text{such that } \Pi(\mathbf{w}, \mathbf{w}') = \{\boldsymbol{\pi} | \boldsymbol{\pi} \geq \mathbf{0}, \boldsymbol{\pi} \mathbf{1}_{n'} = \mathbf{w}, \boldsymbol{\pi}^\top \mathbf{1}_n = \mathbf{w}'\}. \tag{4}$$

With this formulation, Gromov-Wasserstein optimal transport considers aligning a pair of samples $\mathbf{x}_i, \mathbf{x}_k$ in one dataset ($\mathbf{X}$) with a pair of samples $\mathbf{x}'_j, \mathbf{x}'_l$ in the other dataset ($\mathbf{X}'$) by comparing the distances between sample pairs in each domain D^X_{ik} and $D^{X'}_{jl}$. Similarly to 2, the linear constraints placed on the coupling matrix requires that the probability distributions in its column marginals and row marginals match the empirical probabilities defined over the datasets. Observing that this constraint in practice can lead to undesirable alignments for datasets with different representations of cell-type proportions, Pamona [10] and SCOTv2 [11] proposed variations on SCOT to relax this constraint in different ways. Despite these variations, all three methods construct nearest neighbor graphs on the input datasets and compute pairwise distances on these graphs to extract intra-domain sample distances. This procedure discards the features, which are not considered in the alignment.

2.3 Single-cell Co-Optimal Transport (SCOOTR)

Unlike the previous optimal transport-based single-cell alignment methods, SCOOTR does not discard dataset features and optimizes over two coupling matrices, one over the samples ($\boldsymbol{\pi}^s$) and one over the features ($\boldsymbol{\pi}^f$) to attain:

$$\min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}'), \boldsymbol{\pi}^f \in \Pi(\mathbf{v}, \mathbf{v}')} \sum_{i,j,k,l} L(X_{i,k}, X'_{j,l}) \boldsymbol{\pi}^s_{ij} \boldsymbol{\pi}^f_{kl} + \Omega(\boldsymbol{\pi}^s, \boldsymbol{\pi}^f) \tag{5}$$

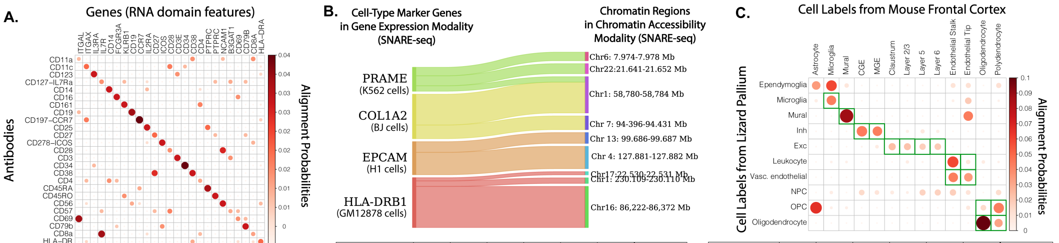
where $\Omega(\boldsymbol{\pi}^s, \boldsymbol{\pi}^f)$ is the entropic regularization term with $\Omega(\boldsymbol{\pi}^s, \boldsymbol{\pi}^f) = \varepsilon_1 H(\boldsymbol{\pi}^s | \mathbf{w}\mathbf{w}'^T) + \varepsilon_2 H(\boldsymbol{\pi}^f | \mathbf{v}\mathbf{v}'^T)$ and $H(\boldsymbol{\pi}^s | \mathbf{w}\mathbf{w}'^T) = \sum_{i,j} \log(\frac{\pi^s_{ij}}{w_i w'_j}) \pi^s_{ij}$ being the relative entropy. Here, $\mathbf{w} \in \Delta_n$ and $\mathbf{w}' \in \Delta'_n$ represent the empirical measures defined over samples, as described in Section 2.1, and similarly, $\mathbf{v} \in \Delta_d$ and $\mathbf{v}' \in \Delta'_d$ are uniform measures defined over the features. This time, while the scripts i and j still refer to sample indices, k and l refer to feature indices in the datasets $\mathbf{X}$ and $\mathbf{X}'$, respectively. Intiutively, $L(X_{i,k}, X'_{j,l}) = (X_{i,k} - X'_{j,l})^2$ compares each feature in each pair of cells across the two modalities after both the cells and the features of one modality have been transformed with respect to the two coupling matrices. Since the feature space is also being transformed by $\boldsymbol{\pi^f}$, this formulation allows us to compare multi-omic datasets without discarding features. The hyperparameters ε_1 and ε_2 control the extent of entropic regularization over the two coupling matrices, which controls their sparsity.

This optimization formulation is based on Co-Optimal Transport, introduced by Redko *et al* [15], which uses and alternating block coordinate descent procedure to solve for both $\boldsymbol{\pi^s}$ and $\boldsymbol{\pi^f}$. We describe the details of the optimization procedure in Supplementary Algorithm 1.

One of the advantages of aligning both the samples and the features is the opportunity to provide supervision on one of them (e.g. cell-type alignments) to improve alignments on the other (e.g. features, or vice-versa). To do this, we optionally provide a "penalization matrix" to scale the cost of certain alignments, as detailed in Supplementary Materials.

3 Results

We apply SCOOTR to three real-world datasets with some ground-truth information to benchmark its alignment performance: (1) a CITE-seq dataset, with gene expression measurements and antibody abundance profiles for ten antibodies from human peripheral blood mononuclear cells [2], (2) a SNARE-seq dataset, with chromatin accessibility and gene expression profiles from a mixture of four cell lines: BJ, H1, K562, and GM12878 [1], and finally (3) a multi-species scRNA-seq dataset with gene expression measurements from mouse prefrontal cortex and bearded lizard pallium [16]. For CITE-seq, SCOOTR yields quality alignments for both cells and features in its unsupervised setting. Figure 2(A) visualizes the feature alignment matrix ($\boldsymbol{\pi}^f$) recovered by SCOOTR, where the rows are antibodies and the columns are the ten genes that encode them, in their respective order. We observe that SCOOTR recovers some correspondence for all antibodies and their encoding gene (along the diagonal), and $\sim 72\%$ of the antibodies were assigned the highest coupling probability with their encoding gene. Figure S1(B) visualizes the cell alignments by projecting cells from the one domain (gene expression) onto the cells of the other domain (antibody abundance) by taking a weighted average of cells in the latter domain according to their coupling probabilities the cells in the former domain, as recovered in ($\boldsymbol{\pi}^s$) (a.k.a. "barycentric projection", also used by the previous optimal transport-based alignment methods [9–11]). We visualize that the cells are correctly aligned with their corresponding cell-types and yield a low alignment error of 0.141 (compared to 0.154 by bindSC), as measured by the commonly used "average fraction of sample closer than true match" metric (FOSCTTM) [7, 9, 11, 12].



Supervision level on cell-types	0%	20%	40%	60%	80%	100%	bindSC*
% Accuracy of feature alignments	31.22	40.91	63.84	72.56	79.94	83.89	61.26

Supervision level on features	0%	20%	40%	60%	80%	100%	bindSC* (supervised)
% Accuracy of cell alignments	29.41	35.30	52.49	70.59	82.35	82.35	76.47

Figure 2: **SCOOTR feature alignment results A.** Feature coupling matrix between the ten antibodies and their encoding genes of the CITE-seq dataset. Larger and darker circles correspond to higher alignment probabilities. **B.** Sankey plot visualizing example feature alignments for the four cell-type marker genes and their strongest chromatin accessibility correspondences in the SNARE-seq dataset. The table below indicates feature alignment accuracy at increased level of supervision on cell-type alignments. **C.** Cell-type coupling matrix when full supervision is provided on paralogous genes between the two species. The green boxes indicate relevant cell-type pairings based on prior knowledge.

We observe that, when aligning datasets with more complex relationships, such as chromatin accessibility and gene expression alignments for the SNARE-seq dataset, where the underlying correspondences are not expect to be $1-1$, supervision on cell-type annotations improves feature alignment performance. Figure 2(B) visualizes an example of feature alignments recovered by SCOOTR for the four cell-type marker genes in this dataset, with validations from the literature described in Supplementary Materials. The table below this figure shows the increase in feature alignment accuracy with supervision, as benchmarked against the regulatory relationships predicted by CellOracle software [17], which contructs gene regulatory networks based on gene expression and chromatin accessibility data. Similarly, we observe that supervision on the feature level improves cell-type alignments for the cross-species gene expression dataset. In Figure 2(C), we visualize the cell-type alignments between the gene expression datasets for the two species, after averaging cell alignment probabilities based on cell-type annotations. Here, we provide full supervision on the feature-level alignments by only penalizing the alignment of non-paralogous gene pairs. As the table below this figure indicates, cell-type alignment improves with increased percentage of the paralogous genes used for supervision. We compare our fully supervised alignment accuracy (82.35%) with the bindSC (76.47%), which is also in its fully supervised setting. Since bindSC requires a prior on feature matchings and this dataset involves the alignment of the same modality (gene expression), we construct this prior matrix ("gene activity matrix") based on paralogous gene matches. Despite this, we find that the cell-type alignments by SCOOTR are more accurate than the alignments by bindSC.

References

[1] Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019.

[2] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

[3] Stephen J. Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle, and Wolf Reik. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature Communications*, 9(1):781, 2018.

[4] Anjun Ma, Adam McDermaid, Jennifer Xu, Yuzhou Chang, and Qin Ma. Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*, 38(9):1007–1022, 2020.

[5] Michael Eisenstein. The secret life of cells. *Nature Methods*, 17(1):7–10, 2020.

[6] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly Embedding Multiple Single-Cell Omics Measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143, pages 10:1–10:13, 2019.

[7] Ritambhara Singh, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, and William Stafford Noble. Unsupervised manifold alignment for single-cell multi-omics data. ACM-BCB, 2020.

[8] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, 2020.

[9] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020.

[10] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, 08 2021. btab594.

[11] Pınar Demetci, Rebecca Santorella, Björn Sandstede, and Ritambhara Singh. Unsupervised Integration of Single-Cell Multi-omics Datasets with Disproportionate Cell-Type Representation. In *26th International Conference on Research in Computational Molecular Biology (RECOMB 2022)*, pages 3–19. Springer International Publishing, 2022.

[12] Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, Li Li, May Daher, Rafet Basar, Katayoun Rezvani, Rui Chen, and Ken Chen. Bi-order multimodal integration of single-cell data. *Genome Biology*, 23(1):112, 2022.

[13] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, September 2008.

[14] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11:355–607, 2019.

[15] Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Co-optimal transport, 2020.

[16] April R. Kriebel and Joshua D. Welch. Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature Communications*, 13(1):780, 2022.

[17] Kenji Kamimoto, Christy M. Hoffmann, and Samantha A. Morris. Celloracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, 2020.