Gromov-Wasserstein Optimal Transport to Align Single-Cell Multi-Omics Data

Pinar Demetci* [12], **Rebecca Santorella*** [3] Björn Sandstede [3], William Stafford Noble [45], and Ritambhara Singh [1,2]

[1]Department of Computer Science, Brown University [2]Center for Computational Molecular Biology, Brown University [3]Division of Applied Mathematics, Brown University
[4]Department of Genome Sciences, University of Washington [5]Paul G. Allen School of Computer Science and Engineering, University of Washington

Data integration is important but challenging

- It is difficult to apply multiple sequencing technologies to the same cell
- Thus, data loses cell-to-cell correspondence across domains
- We need unsupervised alignment algorithms to recover cell-to-cell correspondences

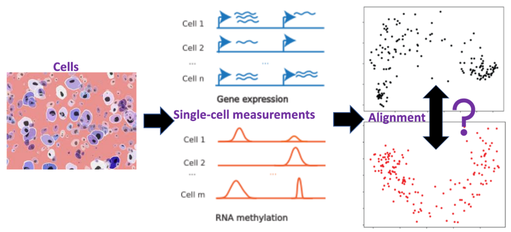

Figure 1: To study genomic heterogeneity, we need to be able to align data sets from single cell measurements without cell-to-cell or feature-to-feature correspondences.

We present SCOT, an unsupervised alignment algorithm that uses Gromov-Wasserstein optimal transport to find a probabilistic mapping between samples from two sequencing domains.

SCOT yields state-of-the-art alignment but in less time and with fewer hyperparameters.

Previous Unsupervised Alignment Algorithms

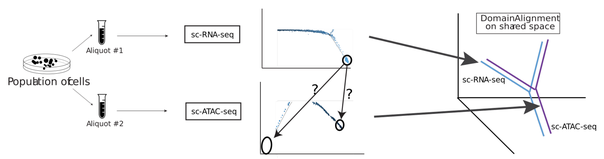

Figure 2: Previous methods attempt to discover the underlying manifold structure

- MMD-MA [4] and UnionCom [1] align and embed the data into a new space
- Both methods require 4 hyperparameters

Discrete optimal transport assigns probabilities between data points in different domains

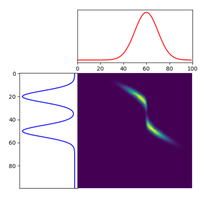

Figure 3: Coupling matrices relate probability distributions

- Optimal transport finds the most cost-effective way to transform one probability distribution into another [5]
- Discrete optimal transport outputs a coupling matrix Γ where each entry Γ_{ij} assigns a probability that sample i in the first domain corresponds to sample j in the second domain
- Gromov-Wasserstein optimal transport preserves intra-domain pairwise distances [6]

Single-Cell alignment using Optimal Transport (SCOT)

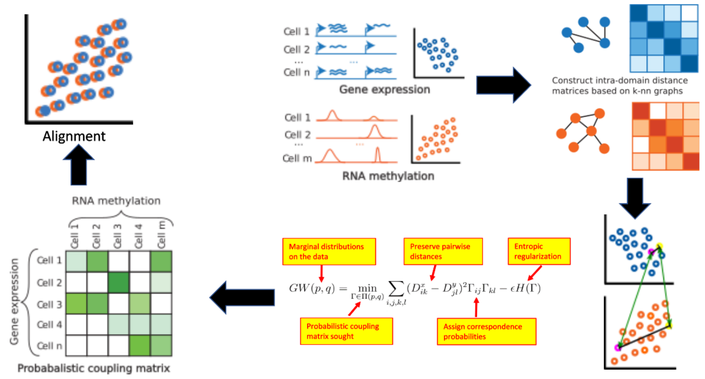

Figure 4: (1) Compute intra-domain distance matrices through k-nearest neighbor (kNN) graphs, (2) Obtain a coupling matrix via entropically regularized Gromov-Wasserstein optimal transport, and (3) Align the data sets via barycentric projection

SCOT gives state-of-the-art performance for single-cell multi-omics alignment

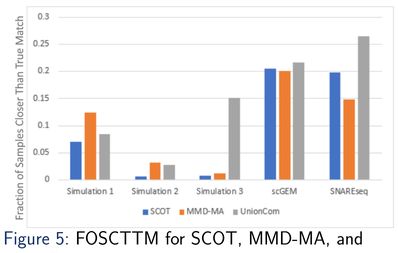

Figure 5: FOSCTTM for SCOT, MMD-MA, and UnionCom

- We compare SCOT to MMD-MA and UnionCom for two real-world data sets.
- scGEM [3] co-assays gene expression and DNA methylation
- SNAREseq [2] co-assays chromatin accessibility and gene expression.

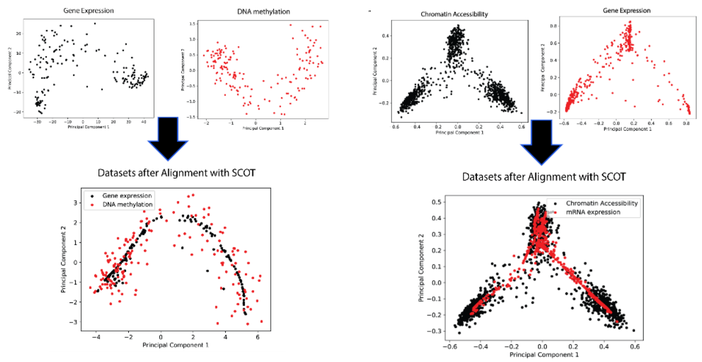

Figure 6: SCOT's alignment for scGEM (left) and SNAREseq (right)

SCOT successfully aligns simulated data

We benchmark SCOT on three simulated data sets from [4]:

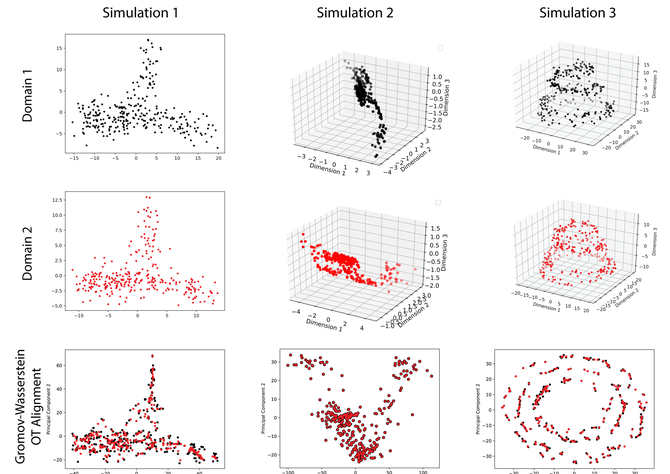

Figure 7: **Top:** PCA projections of domain 1, **Middle:** PCA projections of domain 2, **Bottom:** SCOT's alignment visualized with PCA projections

SCOT is faster than other alignment algorithms and has fewer hyperparameters

Our method, SCOT,

- performs on par with other methods,
- has only 2 hyperparameters, and
- is on average 25 times faster than previous algorithms

Full details: `https://tinyurl.com/SCOT20`

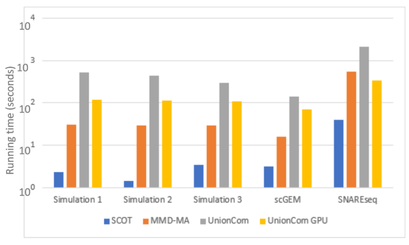

Figure 8: Run times for SCOT, MMD-MA, UnionCom, and UnionCom GPU

Acknowledgements

William S. Noble's contribution to this work was funded by NIH award U54 DK107979. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1644760.

References

[1] Kai Cao et al. "Unsupervised Topological Alignment for Single-Cell Multi-Omics Integration". In: *bioRxiv* (2020).
[2] Song Chen, Blue B Lake, and Kun Zhang. "High-throughput sequencing of transcriptome and chromatin accessibility in the same cell". In: *Nature Biotechnology* 37.12 (2019), pp. 1452–1457.
[3] Lih Feng Cheow et al. "Single-cell multimodal profiling reveals cellular epigenetic heterogeneity". In: *Nature Methods* 13.10 (2016), pp. 833–836.
[4] Jie Liu et al. "Jointly embedding multiple single-cell omics measurements". In: *BioRxiv* (2019), p. 644310.
[5] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.
[6] Gabriel Peyré, Marco Cuturi, and Justin Solomon. "Gromov-wasserstein averaging of kernel and distance matrices". In: *International Conference on Machine Learning*. 2016, pp. 2664–2672.