BINF 3360 FINAL PROJECT: LITERATURE REVIEW
Emily Liau

INTRODUCTION

The following sections provide an overview of single-cell RNA sequencing, dimensionality reduction, and the primary characteristics of the t-SNE and UMAP algorithms.

Single-Cell RNA Sequencing

The following sections detail the characteristics of single-cell RNA sequencing data, experiments, and analysis workflows.

scRNA-seq versus Bulk RNA-seq

Single-cell RNA sequencing (scRNA-seq) refers to the process of sequencing the transcriptomes of individual cells, establishing a high-resolution overview of cell-to-cell variation (Jovic et al., 2022). Compared to bulk RNA-seq, scRNA-seq may reveal the primary cell types and their associated functions within a biological system, allowing individuals to assess cellular heterogeneity across distinct samples in a high-throughput sequencing experiment. Although this sequencing approach provides broader insight into behavior at the cellular level, scRNA-seq experiments are often more expensive and time-consuming compared to bulk RNA-seq.

scRNA-seq Data Analysis

scRNA-seq data from Cell Ranger by 10x Genomics are structured as filtered feature-barcode matrices, with features as rows and barcodes as columns. In a filtered feature-barcode matrix, only cell-associated barcodes are included, and all non-targeted genes are removed during the processing stage. Prior to deriving biological insights, scRNA-seq data are filtered through a multi-step analysis workflow, encompassing quality control, normalization, feature selection, dimensionality reduction, and clustering (Slovin et al., 2021). Overall, the established thresholds and parameters across each stage of the workflow may have a substantial impact on the results obtained from downstream analyses.

Dimensionality Reduction

Dimensionality reduction refers to the process of transforming data from a high-dimensional space into a lower-dimensional space, while preserving the majority of the intrinsic structure and composition of the original dataset. Dimensionality reduction techniques can be categorized into unsupervised/supervised, linear/non-linear, and parametric/non-parametric algorithms. Non-linear dimensionality reduction techniques seek to identify low-dimensional manifolds that accurately represent high data density, to effectively map the data from its high-dimensional representation into a lower-dimensional embedding (DeMers, 1992).

t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique commonly applied towards high-dimensional data, and transforms the data by constructing a Student's t-distribution to calculate the similarity between two points in the low-dimensional embedding (van der Maaten, 2008). Although t-SNE accurately represents the local structure of high-dimensional data, it fails to preserve its global structure, often leading to ambiguous or misleading data visualizations (Kobak, 2019).

Algorithm Parameters

The primary parameters of t-SNE are perplexity, step, and epsilon values. Perplexity refers to the balance between the local and global structure of the data, and is established based on the number of nearest neighbors of a sample within the dataset. According to van der Maaten (2008), optimal perplexity values

range from 5 - 50, and larger parameter values facilitate the preservation of global structure following dimensionality reduction. The step size refers to the number of iterations required to stabilize the resulting structure of the data, and epsilon represents the learning rate, a parameter tuned after gaining initial knowledge of the size of the data. As t-SNE computes the pairwise distances between points, it has a runtime complexity of $O(N^2)$.

Advantages and Limitations

According to van der Maaten (2008), t-SNE accurately reduces high-dimensional data into 2-3 dimensions, enabling the visualization of the relative proximities between clusters in the resulting embedding. However, the resulting insights derived from the visualizations may be misleading, as t-SNE is sensitive to hyperparameters. In addition, variations in the perplexity or seed values may impact the structure presented in the resulting embedding.

t-SNE is less effective on noisy datasets, and does not scale for larger datasets, leading to loss of global information. From an algorithmic standpoint, t-SNE is computationally expensive compared to UMAP and additional dimensionality reduction techniques. Thus, principal component analysis (PCA) often serves as initial pre-processing, in order to extract signal from noise prior to applying t-SNE for further dimensionality reduction.

Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction technique commonly applied towards high-dimensional data, and transforms the data by constructing graph representations such that the low-dimensional and high-dimensional embeddings are structurally similar (McInnes et al., 2020). The UMAP algorithm assumes the Riemannian metric is locally constant, and that the data follows a uniform distribution on a locally-connected Riemannian manifold. Compared to t-SNE, UMAP preserves a larger majority of the global structure of the data while maintaining a reasonable runtime complexity.

Algorithm Parameters

The primary parameters of UMAP are the number of neighbors and minimum distance. According to Armstrong et al. (2021), increasing the number of neighbors increases the number of global distances preserved following dimensionality reduction. The minimum distance between the samples affects the distribution of clustering in the resulting embedding, where lower values represent a greater degree of similarity between dense clusters of samples.

Advantages and Limitations

Compared to t-SNE, UMAP is computationally efficient and preserves a greater amount of the global structure within the 2D projection of the data. Although UMAP does not require pre-processing with PCA and scales appropriately for larger datasets (Nanga et al., 2021), it may not always preserve the structure of data with a more complex topology.

CURRENT APPLICATIONS

The following sections detail the current applications of dimensionality reduction techniques towards various types of biological data, including scRNA-seq and microbiome sequencing data.

Dimensionality Reduction for scRNA-seq Data

Due to low coverage in high-throughput sequencing experiments, dimensionality reduction techniques are commonly applied towards single-cell RNA-sequencing (scRNA-seq) data to overcome the inherent sparsity, variance, and high dimensionality. In addition, dimensionality reduction techniques serve to uphold the balance between the local and global structure of a given dataset (Kobak, 2019). With regards to scRNA-seq, local structure isolates specific cell types as distinct clusters, and allows individuals to

accurately interpret variation in gene expression contributing to cellular heterogeneity within scRNA-seq datasets. In contrast, global structure serves to maintain the inter-cluster embeddings and distances within the dataset. Overall, non-linear dimensionality reduction techniques enhance the interpretability of corresponding data visualizations without compromising the underlying global structure of the data, allowing individuals to observe the underlying clustering patterns and more accurately derive biological insights from a given dataset (Xiang et al., 2021).

Dimensionality Reduction for Microbiome Data

According to Armstrong et al. (2022), microbiome data is characteristic of high dimensionality, sparsity, and compositionality. Compositionality refers to the degree of randomness of the reads in a specific sample, due to random sampling during data collection. Thus, increased randomness within a microbiome sequencing dataset leads to false correlations in resulting insights, and requires normalization and dimensionality reduction techniques to mitigate the degree to which the number of sequences in a certain sample impact the distances between samples. In addition, the objective of microbiome analyses is to identify biomarkers present across multiple microbial species, to facilitate disease diagnosis or quantify microbiome composition. Thus, dimensionality reduction techniques serve to overcome the hierarchical structure of microbiome data, in order to increase interpretability in downstream analyses.

FUTURE PROCEEDINGS

Overall, the resulting visualizations generated in this tutorial are limited by the constraints of the dataset. More specifically, we did not include cell type annotations associated with the clusters in t-SNE and UMAP in the combined dataframe of all samples. In the future, we may consider assessing the clustering patterns of various cell types across the three regions of the small intestine of Mus musculus or other species.

Broadly, dimensionality reduction techniques serve to facilitate interpretable data visualization of feature representations, enhance downstream analyses, mitigate overfitting, and reduce the runtime complexity of machine learning algorithms. From a computational and algorithmic standpoint, researchers may continue applying clustering algorithms such as K-Means clustering or Density-Based Spatial Clustering of Applications with Noise (DBSCAN), in order to assess the impact of various distance metrics and hyperparameters on downstream analyses for scRNA-seq data (Yigin et al., 2023). In addition, researchers may also assess the impact of dimensionality reduction techniques on additional types of biological data across various species, in order to improve interpretability and derive more accurate biological insights.

REFERENCES

Adam L. Haber. (n.d.). A single-cell survey of the small intestinal epithelium | Nature. https://www.nature.com/articles/nature24489

Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., & Knight, R. (2021). Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. MSystems, 6(5), e00691-21. https://doi.org/10.1128/mSystems.00691-21

Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., & Knight, R. (2022). Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. Frontiers in Bioinformatics, 2. https://www.frontiersin.org/articles/10.3389/fbinf.2022.821861

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

DeMers, D., & Cottrell, G. (1992). Non-Linear Dimensionality Reduction. Advances in Neural Information Processing Systems, 5. https://proceedings.neurips.cc/paper/1992/hash/cdc0d6e63aa8e41c89689f54970bb35f-Abstract.html

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology, 20(1), 296. https://doi.org/10.1186/s13059-019-1874-1

Hippen, A. A., Falco, M. M., Weber, L. M., Erkan, E. P., Zhang, K., Doherty, J. A., Vähärautio, A., Greene, C. S., & Hicks, S. C. (2021). miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. PLoS Computational Biology, 17(8), e1009290. https://doi.org/10.1371/journal.pcbi.1009290

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. Clinical and Translational Medicine, 12(3), e694. https://doi.org/10.1002/ctm2.694

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nature Communications, 10(1), Article 1. https://doi.org/10.1038/s41467-019-13056-x

Kobak, D., & Linderman, G. C. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization (p. 2019.12.19.877522). bioRxiv. https://doi.org/10.1101/2019.12.19.877522

Kong, S., Zhang, Y. H., & Zhang, W. (2018). Regulation of Intestinal Epithelial Cells Properties and Functions by Amino Acids. BioMed Research International, 2018, 2819154. https://doi.org/10.1155/2018/2819154

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2014). Laplacian Dynamics and Multiscale Modular Structure in Networks. IEEE Transactions on Network Science and Engineering, 1(2), 76–90. https://doi.org/10.1109/TNSE.2015.2391998

Lytal, N., Ran, D., & An, L. (2020b). Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. Frontiers in Genetics, 11, 41. https://doi.org/10.3389/fgene.2020.00041

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Nanga, S., Bawah, A. T., Acquaye, B. A., Billa, M.-I., Baeta, F. D., Odai, N. A., Obeng, S. K., & Nsiah, A. D. (2021). Review of Dimension Reduction Methods. Journal of Data Analysis and Information Processing, 9(3), Article 3. https://doi.org/10.4236/jdaip.2021.93013

Osorio, D., & Cai, J. J. (2020). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics, 37(7), 963–967. https://doi.org/10.1093/bioinformatics/btaa751

Saini, H., Lal, S. P., Naidu, V. V., Pickering, V. W., Singh, G., Tsunoda, T., & Sharma, A. (2016). Gene masking—A technique to improve accuracy for cancer classification with high dimensionality in microarray data. BMC Medical Genomics, 9(Suppl 3), 74. https://doi.org/10.1186/s12920-016-0233-2

Slovin, S., Carissimo, A., Panariello, F., Grimaldi, A., Bouché, V., Gambardella, G., & Cacchiarelli, D. (2021). Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. Methods in Molecular Biology (Clifton, N.J.), 2284, 343–365. https://doi.org/10.1007/978-1-0716-1307-8_19

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biology, 20(1), 295. https://doi.org/10.1186/s13059-019-1861-6

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. Scientific Reports, 9(1), Article 1. https://doi.org/10.1038/s41598-019-41695-z

van der Maaten, L. (2008, November). Visualizing Data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Vasighizaker, A., Danda, S., & Rueda, L. (2022). Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data. Scientific Reports, 12(1), Article 1. https://doi.org/10.1038/s41598-021-03613-0

Verma, A., & Engelhardt, B. E. (2020). A robust nonlinear low-dimensional manifold for single cell RNA-seq data. BMC Bioinformatics, 21(1), 324. https://doi.org/10.1186/s12859-020-03625-z

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. Genome Biology, 19(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology, 20(1), 59. https://doi.org/10.1186/s13059-019-1663-x

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. Frontiers in Genetics, 12. https://www.frontiersin.org/articles/10.3389/fgene.2021.646936

Yip, S. H., Sham, P. C., & Wang, J. (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Briefings in Bioinformatics, 20(4), 1583–1589. https://doi.org/10.1093/bib/bby011