frontiers | Frontiers in Genetics

EDITORIAL
published: 03 June 2022
doi: 10.3389/fgene.2022.911999

Editorial: Machine Learning and Mathematical Models for Single-Cell Data Analysis

Le Ou-Yang[1], Xiao-Fei Zhang[2], Jiajun Zhang[3], Jin Chen[4] and Min Wu[5]*

[1]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ), College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, [2]School of Mathematics and Statistics and Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, China, [3]Guangdong Province Key Laboratory of Computational Science, School of Mathematics, Sun Yat-sen University, Guangzhou, China, [4]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, United States, [5]Institute for Infocomm Research (I2R), A*STAR, Singapore, Singapore

Keywords: single-cell omics data, machine learning, mathematical modelling, data integration, network modeling

Editorial on the Research Topic

Machine Learning and Mathematical Models for Single-Cell Data Analysis

Understanding how individual cells communicate with each other and respond to evolution and perturbations is a central challenge of biology (Altschuler and Wu, 2010). Due to the heterogeneity of cells, studying a bulk population of cells may confound the variability of cell-type compositions, single cell analysis has the potential to enable a more systematic study of the inner workings of biological systems, and allows us to uncover the underlying mechanisms for cellular functions and biological processes such as cell differentiation and disease development. In the past decade, advances in single-cell isolation and sequencing technologies have enabled the assay of DNA, mRNA, and protein abundances at single-cell resolution, which promote the study of genomics, transcriptomics, proteomics and metabolomics at the sinlge cell level. For example, single-cell genomic analysis can shed light to the genomic variability of individual cells, while single-cell transcriptomic and proteomic analysis can help to reveal the types and functional states of individual cells (Shapiro et al., 2013). However, processing single-cell data of high dimensionality and scale is inherently difficult, especially considering the degree of noise, sparsity, batch effects and heterogeneity in the data (Amodio et al., 2019). Thus, there is an urgent need for developing computational models which can handle the size, dimensionality, and various characteristics of single-cell data. In this Research Topic of Frontiers in Genetics on "*Machine Learning and Mathematical Models for Single-Cell Data Analysis*," we have collected eight manuscripts that used machine learning algorithms or mathematical models to solve problems in single cell analysis.

Single-cell and whole tissue RNA sequencing technologies enable the Research Topic of detailed information about biological processes at genomic and transcriptomic levels. Besides, existing microscopy and cell-resolution imaging techniques allow the high-quality characterization of morphology and physiology at the level of extended fragments of tissues and organs. Bobrovskikh et al. summarized the potential of single-cell technologies together with advanced imaging techniques for computational modelling in plants. They reviewed currently available single-cell data analysis approaches, advanced imaging technologies in plant research with single-cell resolution and cell-based modelling approaches. They shown how the combination of single-cell data, morphometric data and cell-based models help to expand the understanding of tissue and organ morphogenesis.

Tissues are constituted of heterogeneous cell types. Although single-cell RNA sequencing has paved the way to a deeper understanding of organismal cellular composition, the high cost and technical noise have prevented its wide application. As an alternative, computational deconvolution

OPEN ACCESS

Edited and reviewed by:
Alfredo Pulvirenti,
University of Catania, Italy

****Correspondence:***
Le Ou-Yang
leouyang@szu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *03 April 2022*
Accepted: *19 May 2022*
Published: *03 June 2022*

Citation:
Ou-Yang L, Zhang X-F, Zhang J, Chen J and Wu M (2022) Editorial: Machine Learning and Mathematical Models for Single-Cell Data Analysis. Front. Genet. 13:911999. doi: 10.3389/fgene.2022.911999

of bulk tissues can be a cost-effective solution (Jin and Liu, 2021). Liu et al. proposed a deconvolution method, named DecOT, to characterize the cell type composition from bulk tissue RNA-seq data. DecOT uses the optimal transport distance as a loss and applies an ensemble framework to integrate reference information from scRNA-seq data of multiple individuals. Experiment results on real data sets demonstrated that DecOT outperformed other existing methods and was robust to the choice of references.

The development of single-cell sequencing technologies promotes the researches on developmental physiology and disease (Potter, 2018), but the spatial information of individual cells is lost due to the tissue dissociation processes in these technologies. Highly multiplexed imaging technologies, such as imaging mass cytometry (IMC), are powerful tools to exploit the composition and interactions of cells in tumor microenvironments at subcellular resolution. However, due to the high resolution and large number of channels, how to process and interpret IMC image data still remains challenging (Chang et al., 2017). To improve the accuracy of single cell segmentation, which is a critical step to process IMC image data, Xiao et al. developed a deep neural network (DNN)-based cell segmentation method, named Dice-XMBD. Dice-XMBD is marker agnostic and can perform accurate cell segmentation of IMC images of different channel configurations without modification.

Advances in single-cell RNA-sequencing (scRNA-seq) technology provided an unprecedented opportunity for researchers to study the identity and mechanisms of single cells (Morris, 2019). Besides scRNA-seq data, spatial location data can also provide important information on the cells' micro-environment and cell-cell interactions (Mayr et al., 2019), which can contribute to cell type identification. Oh et al. proposed a hybrid clustering approach, named single-cell Hybrid Nonnegative Matrix Factorization (scHybridNMF), to perform cell clustering by jointly processing cell location and gene expression data. ScHybridNMF combines sparse nonnegative matrix factorization (sparse NMF) with k-means clustering to cluster high-dimensional gene expression and low-dimensional location data. Experiment results on simulated and real data sets demonstrate the effectiveness of scHybridNMF in detecting cell clusters.

The communication between cells plays a vital role in the development, physicology, and pathology of muticellular organisms. Single-cell RNA-sequencing (scRNA-seq), which measures the expression levels of a great number of genes across various cell types at single-cell resolution, provides a great opportunity to study the cell-cell communication between interacting cells and the signaling response governed by intracellular gene regulatory networks (GRNs) (Shao et al., 2020). Identification the changes of intercellular signaling across different conditions is crucial for understanding how distinct cell states respond to evolution, perturbations, and diseases. Wang et al. generalized their previously developed tool CellChat to enable a flexible comparison analysis of cell-cell communication networks across multiple conditions, which facilitated the detection of signaling changes of cell-cell communication in response to biological perturbations. By studying the signaling changes across three mouse embryonic developmental stages, four time points after mouse spinal cord injury, and patients with different COVID-19 severities (i.e., control, moderate, and critical cases), they verified the effectiveness of their proposed approaches. To infer the changes of GRNs between two different states, Liu et al. proposed a general differential network inference framework, named weighted joint sparse penalized D-trace model (WJSDM). WJSDM can directly infer the differential network between two different states by integrating multi-platform gene expression data and various existing biological knowledge. By applying WJSDM to the gene expression data of ovarian cancer and the scRNA-seq data of circulating tumor cells of prostate cancer, and infer the differential network associated with platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer, the authors found some important biological insights about the mechanisms underlying platinum resistance of ovarian cancer and anti-androgen resistance of prostate cancer.

Recent advances in experimental biology have generated huge amounts of data. For example, Microwell-Seq, a single-cell RNA-sequencing technology, has been used to analyze the transcriptome of more than 400,000 mouse single cells, covering all major mouse organs (Han et al., 2018). There is an urgent need for next generation methods to deal with large, heterogeneous and complex data sets Camacho et al. (2018). As a promising data processing method, deep learning methods have been employed in biological data processing (Eraslan et al., 2019). However, the deep learning methods usually run as a "black box," which is hard to interpret. The capsule network (CapsNet) is a newly developed deep learning model for digital recognition tasks (Sabour et al., 2017). Wang et al. (2020) proposed a modified CapsNet model, called single cell capsule network (scCapsNet), which is a highly interpretable cell type classifier, with the capability of revealing cell type associated genes by model internal parameters. Based on CapsNet and scCapsNet, Wang et al. proposed a deep learning classifier and data integrator, named MultiCapsNet. The MultiCapsNet model could integrate multiple input sources and standardize the inputs, then use the standardized information for classification through capsule network. The experiment results on three data sets with different data type and application scenarios proved the validity and interpretability of MultiCapsNet.

Cancer immunotherapy has shown to elicit substantial response to many cancers and has led to significant increases in quality of life for cancer patients. This is especially true of checkpoint therapy, which causes tumor regression in previously untreatable cancers. However, the potential mechanisms of checkpoint therapy are still being investigated and there are as of yet few prognostic markers for response (Bai et al., 2020). Immune checkpoint therapies such as PD-1 blockade have vastly improved the treatment of numerous cancers, including basal cell carcinoma (BCC). However, patients afflicted with pancreatic ductal carcinoma (PDAC), one of the deadliest malignancies, overwhelmingly exhibit negative responses to checkpoint therapy. Liu et al. sought to combine data analysis and machine learning to differentiate the putative mechanisms of BCC and PDAC non-response. By comparing two recent single-

cell transcriptomic datasets of PDAC and BCC, the authors identified some potential biomarkers and mechanisms related to BCC and PDAC non-response. By utilizing machine learning classification algorithms, they also discovered that PDAC displays greater similarities to melanoma, which is highly immunogenic and undergoes rapid metastasis, than to BCC (Dollinger et al., 2020).

In summary, this Research Topic covers various aspects of machine learning models, including supervised and unsupervised approaches and their applications for single-cell data analysis, which paves the way for using machine learning and mathematical models in service of various tasks towards single cell analysis. We hope the readers from bioinformatics and the domain specific researchers will be benefitted by reading articles included in this Research Topic.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Altschuler, S. J., and Wu, L. F. (2010). Cellular Heterogeneity: Do Differences Make a Difference? *Cell.* 141, 559–563. doi:10.1016/j.cell.2010.04.033

Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., et al. (2019). Exploring Single-Cell Data with Deep Multitasking Neural Networks. *Nat. Methods* 16, 1139–1145. doi:10.1038/s41592-019-0576-7

Bai, R., Lv, Z., Xu, D., and Cui, J. (2020). Predictive Biomarkers for Cancer Immunotherapy with Immune Checkpoint Inhibitors. *Biomark. Res.* 8, 34–17. doi:10.1186/s40364-020-00209-0

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation Machine Learning for Biological Networks. *Cell.* 173, 1581–1592. doi:10.1016/j.cell.2018.05.015

Chang, Q., Ornatsky, O. I., Siddiqui, I., Loboda, A., Baranov, V. I., and Hedley, D. W. (2017). Imaging Mass Cytometry. *Cytometry* 91, 160–169. doi:10.1002/cyto.a.23053

Dollinger, E., Bergman, D., Zhou, P., Atwood, S. X., and Nie, Q. (2020). Divergent Resistance Mechanisms to Immunotherapy Explain Responses in Different Skin Cancers. *Cancers* 12, 2946. doi:10.3390/cancers12102946

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep Learning: New Computational Modelling Techniques for Genomics. *Nat. Rev. Genet.* 20, 389–403. doi:10.1038/s41576-019-0122-6

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell.* 172, 1091–1107. doi:10.1016/j.cell.2018.02.001

Jin, H., and Liu, Z. (2021). A Benchmark for Rna-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* 22, 102–123. doi:10.1186/s13059-021-02290-6

Mayr, U., Serra, D., and Liberali, P. (2019). Exploring Single Cells in Space and Time during Tissue Development, Homeostasis and Regeneration. *Development* 146, dev176727. doi:10.1242/dev.176727

Morris, S. A. (2019). The Evolving Concept of Cell Identity in the Single Cell Era. *Development* 146, dev169748. doi:10.1242/dev.169748

Potter, S. S. (2018). Single-cell Rna Sequencing for the Study of Development, Physiology and Disease. *Nat. Rev. Nephrol.* 14, 479–492. doi:10.1038/s41581-018-0021-7

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic Routing between Capsules. *Adv. neural Inf. Process. Syst.* 30, 344. doi:10.1097/01.asw.0000521116.18779.7c

Shao, X., Lu, X., Liao, J., Chen, H., and Fan, X. (2020). New Avenues for Systematically Inferring Cell-Cell Communication: through Single-Cell Transcriptomics Data. *Protein Cell.* 11, 866–880. doi:10.1007/s13238-020-00727-5

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science. *Nat. Rev. Genet.* 14, 618–630. doi:10.1038/nrg3542

Wang, L., Nie, R., Yu, Z., Xin, R., Zheng, C., Zhang, Z., et al. (2020). An Interpretable Deep-Learning Architecture of Capsule Networks for Identifying Cell-type Gene Expression Programs from Single-Cell Rna-Sequencing Data. *Nat. Mach. Intell.* 2, 693–703. doi:10.1038/s42256-020-00244-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.