UNIVERSITY of HOUSTON

CENTER FOR ADVANCED COMPUTING & DATA SCIENCE

Deep Learning to Classify Single-Cell RNA Sequencing in Primary Glioblastoma

Pablo Guillen, Melvin Robinson, Jerry Ebalunode

Abstract

Recent advances in single-cell RNA sequencing technologies enable deep insights into cellular development, gene regulation, and phenotypic diversity by measuring gene expression for thousands of cells in a single experiment. This results in high-throughput datasets and requires the development of new types of computational approaches to extract the useful and valuable underlying biological information of individual cells in heterogeneous biological populations. To addresses these approaches, in this work, we introduce a deep learning technique to classify single cell types data from five primary glioblastomas. We show that the deep learning method has the ability to correctly infer and classify cell type not used during the training process of the algorithm. Further, the deep learning method has the ability to identify the predictor variable Aquaporin 4 (AQP4), as the most important to make these predictions. Such computational approaches, as those presented in this study will enable researchers to better characterize the intratumoral heterogeneity in primary glioblastoma.

Single-cell RNA-sequencing (scRNA-seq)

- scRNA-seq profiles the transcriptome of individual cells
- Heterogeneity within a population of cells
- The identification of new markers for specific types of cells

Glioblastoma

- Glioblastoma is a primary malignant brain tumor developed from star-shaped cells, called astrocytes that support nerve cells
- Glioblastoma, is an archetypal example of a heterogeneous cancer and one of the most lethal human malignancies
- The relationships between different sources of intratumoral heterogeneity: genetic, transcriptional and functional, remain under research

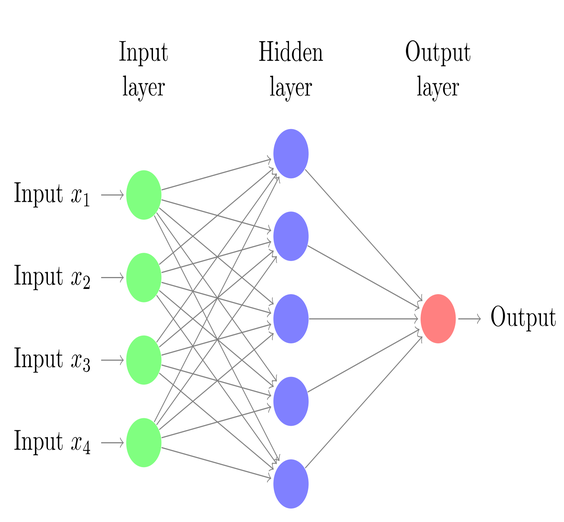
Data Collection

- 430 single gliobastomas cells isolated from 5 five individual tumors
- Data to be procesed contains 5948 rows (genes) quantified in 430 samples (columns)
- Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE57872

Deep Learning

- Deep Learning (DL) is a subfield of machine learning based on learning multiple levels of representations
- DL structure extends the traditional neural networks by adding more hidden layers to the network architecture

Supervised Classification of Single Cells



- Library H2O: Deep Neural Network
- Hidden: (250, 250, 250)
- Optimization method: Adadelta
- Batch size: 20 samples
- Training Epochs: 1000
- Loss function: Cross-entropy
- Activation Function: ReLU
- Weight initialization: normal distribution with mean 0 and std of 0.01

Results

- We trained a DNN algorithm on a set of randomly selected samples, approximately 80% of the entire dataset was used for training, and approximately 20% was used as the testing set

Deep learning
** Reported on training data **
MSE: 5.7716e-21

Deep learning
** Reported on testing data **
MSE: 0.02

- We trained a DNN algorithm using 3-fold cross validation technique

Deep learning
Accurracy: 0.988
MSE-Cross-validation: 0.029

TABLE I. VARAIBLE IMPORTANCES SCORES

Variable	Relative Importance
AQP4	1.00
CADPS	0.98
SGK1	0.97
AXL	0.97
DPP6	0.96
NUDT4	0.95
CRB1	0.95
IGDCC4	0.95
JAG1	0.95
ARHGAP26	0.94

Discussion

- In this study we proposed an efficient methodology which combine gene expression as features with the deep neural network to classify 5 types of primary tumors
- Using the deep neural network classifier shows high accuracy when a discrimination between the classes is executed
- The machine learning method used in this study was able to identifty the most important gene - AQP4 which has been identifed experimentally to play a significcant role in glioma malignancies
- The good results achieved using this computational approach could be employed to evaluate the relationships between different sources of intratumoral heterogeneity in glioblastomas
- We conclude that machine learning techniques, like deep learning, in combination with new molecular techniques, hold promise for improving diagnosis, better assessment of recurrence risk, careful selection of therapy and identification of targets involved in carcinogenesis and function of tumors cells

References

[1] Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O: Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. Nature biotechnology 2015, 33(2):155–160

[2] Patel Anoop P, Tirosh Itay, Trombetta John J, Shalek Alex K, Gillespie Shawn M, Wakimoto Hiroaki, Cahill Daniel P, Nahed Brian V, Curry William T, Martuza Robert L,: Single- cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014, 344(6190):1396 – 1401

[3] Yu-Long Lan, Xun Wang, Jia-Cheng Lou, Xiao-Chi Ma, Bo Zhang: The potential roles of aquaporin 4 in malignant gliomas. Oncotarget, 2017, Vol. 8, (No. 19), pp: 32345-32355

[4] Y. LeCun, Y. Bengio, G. Hinton, “Deep Learning,” Nature 521, pp. 436–444, 2015

[5] S. Aiello, C. Click, H. Roark, L. Rehak, “Machine Learning with Python and H20,” Edited by Lanford, J., Published by H20, 2016