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

Pablo Guillen Rondon
Center for Advanced Computing and Data Science
University of Houston
Houston, TX, USA
pgrondon@central.uh.edu

Jerry Ebalunode
Center for Advanced Computing and Data Science
University of Houston
Houston, TX, USA
jebalunode@uh.edu

***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 paper, 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.**

Keywords—single cell, deep learning, glioblastomas

I. Introduction

Single-cell RNA-sequencing (scRNA-seq) has become increasingly popular for profiling transcriptomic states of individual cells in heterogeneous biological populations [1-5]. scRNA-seq which profiles the transcriptome of individual cells (as opposed to ensemble of cells) has already led to several new and interesting findings. These include the level of heterogeneity within a population of cells [6], the identification of new markers for specific types of cells [7], and the temporal stages involved in the progression of various developmental processes [8]. It has been used to profile diverse systems including cancer tumors [9-10], and cell types within the mouse brain [11], amongst others.

Tumor heterogeneity poses a major challenge to cancer diagnosis and treatment. It can manifest as variability between tumors, wherein different stages, genetic lesions or expression programs are associated with distinct outcomes or therapeutic responses [12-14]. Alternatively, cells from the same tumor may harbor different mutations or exhibit distinct phenotypic or epigenetic states [15-18]. Such intratumoral heterogeneity is increasingly appreciated as a determinant of treatment failure and disease recurrence [19].

The brain is one of the most complex organs in the human body that works with billions of cells. A brain tumor arise when there is uncontrolled division of cells forming an abnormal group of cells around or inside the brain. Brain tumors can be classified to benign or malignant. Malignant tumors are cancerous and they could originate from the brain itself (in which case they are called primary malignant tumor) or they could originate from elsewhere in the body and spread to the brain (in which case they are called secondary malignant tumor) [20]. 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 [21]. Intratumoral heterogeneity and redundant signaling routes likely underlie the inability of conventional and targeted therapies to achieve long-term remissions [22-24]. DNA and RNA profiles of bulk tumors have enabled genetic and transcriptional classification of glioblastomas. However, the relationships between different sources of intratumoral heterogeneity: genetic, transcriptional and functional, remain under research. Inter-patient variation and molecular diversity of neoplasic cells within individual glioblastoma has been previously described [10], showing that established glioblastoma subtype classifiers are variably expressed across individual cells within a tumor and demonstrate the potential prognostic implications of such intratumoral heterogeneity.

Deep Learning (DL) is a subfield of machine learning based on learning multiple levels of representations by making a hierarchy of features where the higher levels are defined from the lower levels. DL structure extends the traditional neural networks by adding more hidden layers to the network architecture between the input and output layers to model more complex and nonlinear relationship [25].

The contribution of this paper is applying the deep learning concept to perform an automated single cells classification using a dataset from 430 cells from five primary glioblastomas [10], and measure its performance.

The structure of this paper is organized as follows: Section II described the steps of the materials and methods, section III presents the experimental results and discussion and the conclusion and future work is given in section IV.

II. Materials and Methods

A. Deep Learning

The concept of deep learning originated from artificial neural network research. A multilayer perceptron with many hidden layers is a good example of the models with deep architectures. Deep learning techniques have been applied to a wide variety of problems in recent years [26-27]. In many of these applications, algorithms based on deep learning have surpassed the previous state-of-art performance. At the heart of all deep learning algorithms is the domain independent idea of using hierarchical layers of learned abstraction to efficiently accomplish high-level task. There are several theoretical frameworks for deep learning, and here we summarize the feedforward architecture used by H2O [28]. Multilayer perceptron (MLP) are feed-forward neural networks with architecture composed of the input layer, the hidden layer and the output layer. Each layer is formed from small units known as neurons. Neurons in the input layer receive the input data X and distribute them forward to the rest of the network. In the next layers, each neuron receives a signal, which is a weighted sum of the outputs of the nodes in the previous layer. Inside each neuron, an activation function is used to control the input. Such a network determines a non-linear mapping from an input vector to the output vector, parameterized by a set of network weights, which are referred to as the vector of weights W. The first step in approximating the weight parameters of the model is finding the appropriate architecture of the MLP, where the architecture is characterized by the number of hidden units, the type of activation function, as well as the number of input and output variables. The second step estimates the weight parameters using the training set. Training estimates the weight vector W to ensure that the output is as close to the target vector as possible. The structure of a MLP network is shown in "Fig. 1".

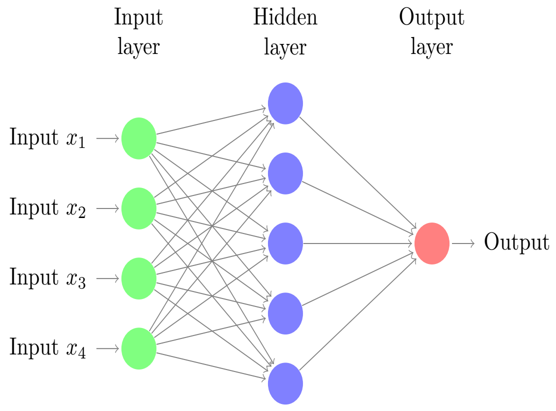

Fig. 1. Structure of an architecture multilayer perceptron

B. Datasets

The dataset consists of 430 single gliobastomas cells isolated from 5 five individual tumors [10]. The matrix data to be procesed contains 5948 rows (genes) quantified in 430 samples (columns). This database has been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE57872.

C. Supervised Classification of Single Cells

In order to apply the deep learning methodology for the classification of single cells a dataset for training and testing with classes was used. The dataset contains 5 classes, where each class considers all the singles cells from a primary tumor.

D. Tools

Library H2O [28] was used in order to perform the classification through deep learning. H2O deep learning is based on a multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation. A feedforward artificial neural network (ANN) model, also known as deep neural network (DNN) or multi-layer perceptron (MLP), is the most common type of Deep Neural Network and the only type that is supported natively in H2O.

In this study we trained a DNN using Anaconda Navigator v1.8.7 y Python 2.7. We stopped the training process after stabilization of the validation accuracy with equal weight for all the classes (1000 epochs). The batch size used is 20 samples. The network weights are initialized randomly, and the ADADELTA adaptive learning rate algorithm is used for weight updates with defaults parameters. The selected loss function is the categorical cross entropy.

III. Experimental Results and Discussion

The experimental took place using two strategies in order to build the predicitve model and evaluate the accurracy of the DNN algorithm:

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

Table I shows the size of the DNN architecture and parameters used on the experiments to evaluate the performance of the classification. Using these parameters allowed achieve higher classification accuracy. ReLU is the non-linear activation function, epochs correspond to the numbers of passes over the training dataset, and nfolds correspond to cross-validation.

TABLE I. PARAMETERS OF THE DNN ARCHITECTURE

Variables	Parameters
input	430
hidden	(250,250,250)
output	5
Activation function	ReLU
Loss function	Cross-entropy
Epochs	1000
nfolds	3

Fig. 2. Parameters of the deep neural network architecture

The evaluation of the performance for the proposed methodology was measured in terms of average classification rate and average area under the ROC curve (AUC) of all the 5 classes (Tumor 1 - class 0, Tumor 2 – class 1, Tumor 3 – class 2, Tumor 4 – class3, and Tumor 5 – class 4).

Table II and Table III shows, using strategy I, the mean square error obtained for training data and testing data, respectively.

TABLE II. PRECISION REPORTED - TRAIN DATA

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

Fig. 3. Mean square error for training data

TABLE III. PRECISION REPORTED - TESTING DATA

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

Fig. 4. Mean square error for testing data

Table IV shows the results, using the strategy II, of the accuraccy and mean square error obtained during the classification using 3-fold-cross validation.

TABLE IV. ACCURRACY

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

Fig. 5. Mean square error for testing data

"Fig. 2". shows the results of the average area under the ROC curve (AUC) of all the 5 classes. We can see that most of the curves follows the left-hand border and then the top border of the ROC space, showing the predictive model has high precision.

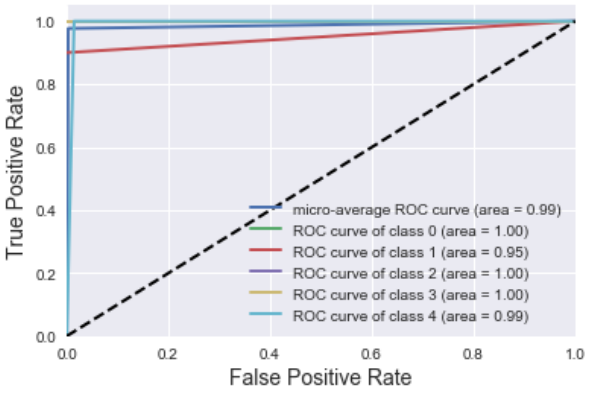

Fig. 6. Area under the ROC curve

H2O deep learning framework has implemented the option to compute variable importances scores and its capable of returning the relative variable importances scores in descending order of importance. Table V shows the results (first 10 variables) of variable importances score for the strategy II, we can observe that DNN is able to indentify which predictor variables are the most important to make these predictions.

TABLE V. 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

Fig. 7. Variable importances scores for the predicitve model

We can see in the top of the Table V the predictor variable AQP4. Recently, accumulated evidence has pointed to AQP4 as a key gene that could play a critical role in glioma development [29].

IV. CONCLUSIONS

In this paper 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 gioblastomas.

Acknowledgment

The authors would like to acknowledge CACDS for its support to present this work. This work was also supported by the National Science Foundation (DMR 1608151).

References

[1] Zheng Grace XY, Terry Jessica M, Belgrader Phillip, Ryvkin Paul, Bent Zachary W, Wilson Ryan, Ziraldo Solongo B, Wheeler Tobias D, McDermott Geoff P, Zhu Junjie, et al.: Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017, 8:14049.

[2] Kowalczyk Monika S, Tirosh Itay, Heckl Dirk, Rao Tata Nages- wara, Dixit Atray, Haas Brian J, Schneider Rebekka K, Wagers Amy J, Ebert Benjamin L, Regev Aviv: Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res 2015, 25(12):1860–1872.

[3] Philipp Angerer, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, David Fischera, Fabian J. Theis: Single cells make big data: New challenges and opportunities in transcriptomics. Current Opinion in Systems Biology 2017, 4:85–91.

[4] Villani Alexandra-Chloé, Satija Rahul, Reynolds Gary, Sarkizova Siranush, Shekhar Karthik, Fletcher James, Griesbeck Morgane, Butler Andrew, Zheng Shiwei, Lazo Suzan: Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 2017, 356(6335):eaah4573.

[5] Aleksandra A. Kolodziejczyk,Jong Kyoung Kim,Valentine Svensson, John C. Marioni, Sarah A. Teichmann: The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell 2015, 58:610-620.

[6] 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.

[7] Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A: Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. Science 2014, 343(6172):776–779. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology 2014, 32(4):381–386.

[8] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L.: The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology 2014, 32(4):381–386.

[9] Chung Woosung, Hyeon Eum Hye, Lee Hae-Ock, Lee Kyung- Min, Lee Han-Byoel, Kim Kyu-Tae, Suk Ryu Han, Kim Sangmin, Eon Lee Jeong, Hee Park Yeon.: Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017, 8.

[10] 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.

[11] Zeisel Amit, Muñoz-Manchado Ana B, Codeluppi Simone, Lönnerberg Peter, Manno Gioele La, Juréus Anna, Marques Sueli, Munguba Hermany, He Liqun, Betsholtz Christer, et al.: Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015, 347(6226): 1138 – 1142.

[12] Yachida S, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467:1114–1117. [PubMed: 20981102]

[13] 2. Eppert K, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med. 2011; 17:1086–1093. [PubMed: 21873988]

[14] 3. Parsons DW, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008; 321:1807–1812. [PubMed: 18772396]

[15] 4. Navin N, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472:90–94. [PubMed: 21399628]

[16] 5. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]

[17] 6. Driessens G, et al. Defining the mode of tumour growth by clonal analysis. Nature. 2012; 488:527– 530. [PubMed: 22854777]

[18] 7. Schepers AG, et al. Lineage Tracing Reveals Lgr5+ Stem Cell Activity in Mouse Intestinal Adenomas. Science. 2012; 337:730–735. [PubMed: 22855427]

[19] 8. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013; 501:355–364. [PubMed: 24048068]

[20] K. Khambhata, S. Panchal, “Multiclass classification of brain tumor in MRI images”, Int J Innov Res Comput Commun Eng. 4 (5) (2016), pp. 8982-8992

[21] Stupp R, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005; 352:987–996. [PubMed: 15758009]

[22] Stommel JM, et al. Coactivation of Receptor Tyrosine Kinases Affects the Response of Tumor Cells to Targeted Therapies. Science. 2007; 318:287–290. [PubMed: 17872411]

[23] Nathanson DA, et al. Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. Science. 2013 10.1126/science.1241328.

[24] Gilbert MR, et al. A Randomized Trial of Bevacizumab for Newly Diagnosed Glioblastoma. New England Journal of Medicine. 2014; 370:699–708. [PubMed: 24552317]

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

[26] M. Langkvist, L. Karlsson, A. Loutfi, “A review of unsupervised feature learning and deep learning for time series modeling,” Pattern Recognition Letters, 42, pp. 11-24, 2014.

[27] D. Yu, L. Deng, “Deep Learning and Its Applications to Signal and Information Processing,” IEEE SIGNAL PROCESSING MAGAZINE, pp. 45-54, 2011.

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

[29] 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.