Gene expression

BRAEM: Basic RNASeq Analysis Employing Machine-Learning

Braden M. H. Katzman[1] and Emily G. Berghoff[1]

[1]Computer Science Department, Columbia University, 116th St and Broadway, New York, NY 10027, USA

ABSTRACT

BRAEM is a program for analyzing single-cell RNA sequencing data and determining cell classification using supervised classification algorithms. The program integrates into one system a variety of popular supervised classification algorithms, including support vector machine, multilayer perceptron (neural network), k-nearest neighbor, and random forest. This flexibility makes BRAEM a useful tool in single-cell RNA sequencing data analysis and classification.

Availability and implementation:

Contact: braden.katzman@columbia.edu; eb2800@columbia.edu

1 BACKGROUND

Determining cell type is a fundamental task in biological research. Not only does cell classification serve to characterize experimental results, but it also has clinical implications, such as in cancer and tumor identification. Better methods for cell classification may allow for novel treatment options by way of personalized medicine.

Cells are commonly identified using features such as location, morphology, and molecular markers. While these methods are reliable for some cell types, they are not adequate for others.

Recent research has successfully demonstrated the utility of single-cell RNA sequencing data in cell identification and classification (Jaitlin *et al.*, 2014; Macosko *et al.,* 2015; Pollen *et al.,* 2014; Treutlin *et al.,* 2014; Zeisel *et al.,* 2015). For example, single-cell RNA sequencing has allowed for the discovery of novel cell types in various tissues and diseases states.

Currently, these data are being analyzed using unsupervised clustering methods. In an effort to expand cell classification algorithms, we have developed BRAEM, a fast and flexible supervised classification program. BRAEM learns the optimal algorithm for single-cell RNA sequencing data classification using limited training data. BRAEM's success demonstrates the ability to employ supervised learning when ground truth data is available and unsupervised methods prove ineffective.

2 IMPLEMENTATION

BRAEM is a python program that leverages the scikit-learn machine learning library (http://scikit-learn.org/).

We tested our program on a single-cell RNA sequencing data set from two types of mouse brain tissue (Zeisel *et al*., 2015), composed of 3005 cells and 19968 genes from 9 classes. This data set allowed us to assess each algorithm's performance on known cell classes.

We preprocessed our data by first removing all genes that had less than 25 molecules across all cells. We then downsampled the number of molecules in each cell to equal the number of molecules in the smallest cell. This removes any technical biases introduced into the data from different sequencing runs while also reducing complexity (Grun *et al*., 2015).

We then processed our data using four supervised learning algorithms: support vector machine, k-nearest neighbor, random forest, and neural network (Kashyap et al., 2014; Libbrecht et al., 2015). We modified scikit's available parameter options using an iterative approach. The results presented reflect the algorithm configurations that converged on maximum accuracy. For support vector machines, we used a radial basis function (rbf) kernel (Scialdone *et al.,* 2015), which suggests non-linear relationships among the cell classes. For

k-nearest neighbor we varied "k" from 1-8, with 2 having the best Matthew's Correlation Coefficient (MCC), a classical metric for machine-learning classification performance (Jurman *et al.*, 2010). For random forest, we chose to set the number of trees i.e. estimators, at ten. For neural networks, we used two hidden layers with a rectified linear unit function for activation, and a stochastic gradient-descent algorithm for backpropagation.

The four algorithms were validated by 10-fold cross validation. In order to compare the algorithms, we employed various evaluation metrics: accuracy, sensitivity, specificity, MCC, and F-score (Jagga *et al.,* 2014; Scialdone *et al.,* 2015).

3 DISCUSSION

One of the current challenges that researchers face with the advancement of sequencing technologies is an overflow of data. BRAEM attempts to address this challenge through the development of more efficient tools for sequence analysis and classification. In addition, BRAEM provides an alternative approach to classification.

Figure 1 illustrates the four machine learning algorithms that BRAEM supports. For our data set, the neural network was most successful at cell classification, as indicated by an average MCC = 0.69. This is relatively good for the field, as other attempts at supervised cell classification have led to MCC values from 0.6-0.9 (Jagga *et al.*, 2014; Scialdone *et al.*, 2015; Hu *et al.,* 2016). Neural networks leverage feature maps generated through data transformations that may prove more effective for classification than methods employed by unsupervised algorithms such as density estimation and k-means clustering. We anticipate that neural network classification will be applicable to other sequencing data sets, such as those used in cancer diagnostics.

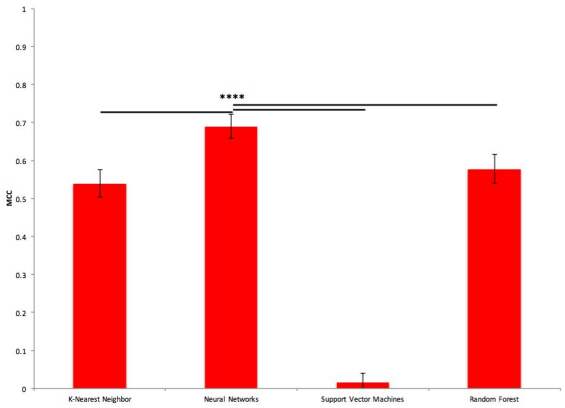

Fig. 1. Comparison of the machine learning algorithms. The average Matthew's Correlation Coefficient (MCC) and the SE are shown. The significance level from a t-test is indicated with "****" meaning $p < 0.0001$.

4 CONCLUSIONS

We have developed a tool, BRAEM, that addresses one of the fundamental data analysis tasks for single-cell RNA sequencing data: cell classification. We demonstrate that this tool can classify a single-cell RNA sequencing data set. We hope that such a classification tool will be robust enough for non-experts to carry out cell classification analysis. To improve upon the success of BRAEM, future work is required to design machines that can search for more complex and hidden features for classification. Convolutional neural networks may prove effective for this task.

Currently, we have implemented our program on one data set (Zeisel *et al.*, 2015). Use of other data sets is planned. Feedback from potential users in that context is highly welcome.

AVAILABILITY AND REQUIREMENTS

BRAEM was written in python and requires python 2.7 and the scikit-learn library (which itself requires NumPy, SciPy and matplotlib).
Source code is freely available at
https://github.com/bradenkatzman/CellClassificationMachineLearning
Scikit-learn is available at
http://scikit-learn.org/stable/

ACKNOWLEDGEMENTS

We thank Dr. Itsik Pe'er for helpful advice. We also thank Dr. Sten Linnarsson for helping us access his previously published single-cell RNA sequencing data.

Funding: none.

Conflict of Interest: none declared.

REFERENCES

Grun,D. *et al.* (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251-255.

Grun,D. and van Oudenaarden,A. (2015) Design and analysis of single-cell sequencing experiments. *Cell* **163** 799-810.

Hu,Y. *et al.* (2016) A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics* **17**, 1025.

Jagga,Z. and Gupta,D. (2014) Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC Proceedings* **8**.

Jaitlin,D.A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779.

Jurman, G. and Furlanello, C. (2010) A unifying view for performance measures in multi-class prediction. https://arXiv:1008.2908 , 1.

Kashyap,H. *et al.* (2014) Big data analytics in bioinformatics: a machine learning perspective. *Journal of Latex Class FIles* **13**.

Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16, 321-332 (2015).

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214.

Pollen,A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology* **32**, 1053-1058.

Scialdone,A. *et al.* (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54-61.

Treutlin,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375.

Zeisel,A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142.