BIOINFORMATICS APPLICATIONS NOTE Vol. 32 no. 1 2016, pages 1-3

Gene expression

BRAEM: supervised classification of single-cell RNA sequencing data

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

Summary: 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 systems, ranging from support vector machines to random forests. This flexibility makes BRAEM a useful tool in single-cell RNA sequencing data analysis and classification.

Availability and implementation: BRAEM was written in python. Source code is freely available at https://github.com/bradenkatzman/CellClassificationMachineLearning

Contact: bmk2137@columbia.edu; eb2800@columbia.edu

Received on May 9, 2016

1 INTRODUCTION

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 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). Currently, these data are being analyzed using unsupervised clustering methods.

In an effort to make cell classification more robust, we have developed BRAEM, a fast and flexible program for cell classification of single-cell RNA sequencing data using supervised classification methods.

2 METHODS

BRAEM is a python program that is built on the scikit (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 with already 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 machines, k-nearest neighbor, random forest, and neural networks (Kashyap et al., 2014; Libbrecht et al., 2015). To implement the supervised learning algorithms, we chose the following parameters. For support vector machines, we used a radial basis function (rbf) kernel (Scialdone *et al.,* 2015). For k-nearest neighbor we varied "k" from 1-8, with 2 having the best Matthew's Correlation Coefficient (MCC). No parameters were needed for random forest and neural networks.

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).

To use the program, the user needs to provide a single-cell RNA sequencing data set in which the classes are already known. The data is then preprocessed, as described above.

After the data has been preprocessed, the four machine learning algorithms are executed. If desired, the user may choose "k = 1-8" for the k-nearest neighbor algorithm. The output consists of text files, labeled according to the algorithm used. The text files contain evaluation metrics for each cell class, as described above.

3 DISCUSSION

Figure 1 illustrates the four machine learning algorithms that BRAEM supports. For our data set, neural networks was most successful at cell classification, as indicated by a MCC = 0.69. This is very good for the field, as other attempts at supervised cell classification have led to MCC values from 0.6-0.8 (Jagga *et al*., 2014; Scialdone *et al*., 2015).

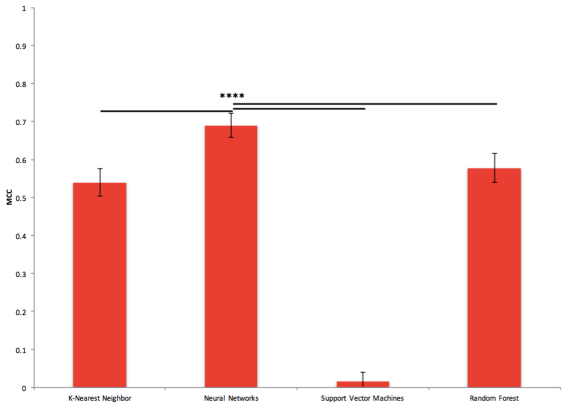

Fig. 1. Comparison of the machine learning algorithms. The average Matthew's Correlation Coefficient (MCC) and the SE are shown. The significance level from a t-test is indicated with "****" meaning $p < 0.0001$.

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. To the authors' knowledge, BRAEM is unique in that it uses supervised learning for cell classification, unlike other methods that rely entirely on unsupervised classification. We hope that such a classification tool will be robust enough for non-experts to carry out cell classification analysis.

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.

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 (2015).

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.

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.