Supervised Learning Approaches for Gut Cell Annotation

Margaret Li

- **Main**: evaluate the performance of supervised learning methods on gut cell datasets
- Cell type annotation based on single-cell RNA-seq data is a critical step for further research on disease and medicine, and supervised learning methods have shown great promise.
- Gut diseases are rising in prevalence (Hills et al), and gut tissues are typically understudied in cell annotation (Svensson et al).
- Evaluating supervised techniques on gut cell annotation would help elucidate how well they generalize to this kind of data

Background

> **Computational cell type annotation methods** typically fall into two categories**:**

1. **Unsupervised cell clustering** to group cells with similar profiles then match them with cell type markers
2. **Supervised learning** to learn from annotated data how to predict cell types on new data

> **Supervised methods** have the advantages of better performance on small or imbalanced data and better scaling. They have grown popular as the amount of high-quality annotated scRNA-seq data has increased. However, their performance is subject to factors like **model construction** and **dataset choice**.

> The **Space-Time Gut Cell Atlas** (Elmentaite et al) is a catalogue of gut cells across space and time: 428,000 cells from the small intestine to the large intestine, and the associated lymph nodes from in utero development to adulthood. I analyze these high quality and comprehensive annotated datasets in my project.

Methods

Train supervised learning models:

1. Random forest
2. SVM with linear kernel
3. SVM with radial basis function kernel
4. Naive Bayes
5. Logistic regression

Evaluate on Space-Time Gut Cell Atlas datasets:

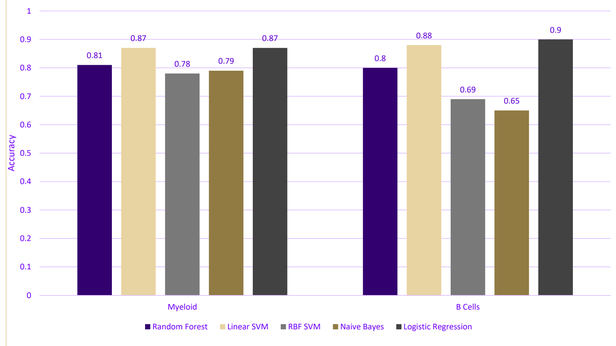
1. Myeloid
2. B Cells

> For each dataset, I set apart 20% of the data as test data, and train on the rest. I extract gene expression data for features, and predict on cell type annotation labels.

> Experiments by Ma et al showed promise for random forest classifier, SVM with linear kernel, and SVM with radial function basis kernel (classifiers 1-3)

> Naive Bayes (4) and logistic regression (5) can be surprisingly powerful on certain datasets if their assumptions hold generally well

Fig. 1: Accuracy of classifiers on myeloid and B Cells data



> As shown in Figure 1 above, logistic regression and SVM with linear kernel outperformed other models, suggesting that a linearity assumption worked well for the data. Consistent with Ma et al, linear SVM had high performance for cell annotation

Results

Performance varied depending on the dataset chosen, and the model used.

> As shown in the below figures, SVM with radial function basis kernel performed generally well on the Myeloid cells, but not so much for the B Cells, and it also had 0 recall and even 0 precision for some underrepresented cell types.

> However, it's possible that further tuning for other models, such as finding good parameters for the SVM with radial basis function kernel could help it achieve better results, and that they could do better on different datasets.

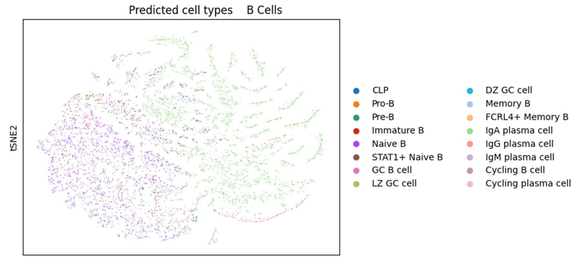
> As shown in the plots to the right, predicted cell types can be represented in a 2D space using t-SNE and seem to cluster.

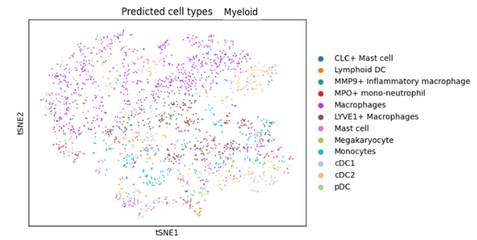
RBF SVM: B Cells

	precision	recall	f1-score
CLP	0	0	0
Cycling B cell	0.57	0.55	0.56
Cycling plasma cell	0.71	0.052	0.097
DZ GC cell	0	0	0
FCRL4+ Memory B	0	0	0
GC B cell	0	0	0
IgA plasma cell	0.91	0.99	0.94
IgG plasma cell	0.89	0.7	0.78
IgM plasma cell	1	0.038	0.073
Immature B	0.71	0.18	0.28
LZ GC cell	0	0	0
Memory B	0.6	0.54	0.57
Naive B	0.41	0.87	0.56
Pre-B	0	0	0
Pro-B	0.67	0.038	0.073
STAT1+ Naive B	0	0	0
accuracy	0.69	0.69	0.69
macro avg	0.4	0.25	0.25
weighted avg	0.65	0.69	0.64

RBF SVM: Myeloid

	precision	recall	f1-score
CLC+ Mast cell	0.88	0.5	0.64
LYVE1+ Macrophages	0.94	0.84	0.89
Lymphoid DC	0.92	0.59	0.72
MMP9+ Inflammatory macrophage	0.87	0.39	0.54
MPO+ mono-neutrophil	0.93	0.57	0.7
Macrophages	0.71	0.94	0.81
Mast cell	0.85	0.91	0.88
Megakaryocyte	1	0.67	0.8
Monocytes	0.87	0.77	0.81
cDC1	0.96	0.5	0.66
cDC2	0.71	0.6	0.65
pDC	0.97	0.68	0.8
accuracy	0.78	0.78	0.78
macro avg	0.88	0.66	0.74
weighted avg	0.8	0.78	0.78





Conclusions

Supervised learning approaches for cell type annotation have deservedly received much attention in recent years. The abilities of supervised learning approaches to learn patterns in annotated single-cell RNA-seq data and then predict the types of cells in new data is extremely promising. I found that performance can vary greatly depending on the dataset and on the model choice, so picking a good model for the data being evaluated is important. Further work to evaluate the abilities of supervised learning methods on diverse tissue types like gut cells is important to advance medicine across the whole of human biology.

References:

1. Hills RD Jr, Pontefract BA, Mishcon HR, Black CA, Sutton SC, Theberge CR. Gut Microbiome: Profound Implications for Diet and Disease. Nutrients. 2019 Jul 16;11(7):1613. doi: 10.3390/nu11071613. PMID: 31315227; PMCID: PMC6682904.
2. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database (Oxford). 2020 Nov 28;2020:baaa073. doi: 10.1093/database/baaa073. PMID: 33247933; PMCID: PMC7698659.
3. Elmentaite, R., Kumasaka, N., Roberts, K. et al. Cells of the human intestinal tract mapped across space and time. Nature 597, 250–255 (2021). https://doi.org/10.1038/s41586-021-03852-1
4. Ma, W., Su, K. & Wu, H. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. Genome Biol 22, 264 (2021). https://doi.org/10.1186/s13059-021-02480-2