Loyola University Chicago - COMP 379/479
Final Project

Using gene expression to classify RNA-Seq samples into tissues

Daniel Araujo, Olaf Garcia, Henry Wittich

1. Introduction

DNA is the biological molecule that contains the blueprint for life in the form of genes. Genes are sequences of nucleotides that serve as a code for creating proteins. The prevalence of different proteins in a cell results in different biological outcomes that account for the large-scale differences between all living organisms. The differences between individuals of different species, individuals of the same species, and even cells within the same individual can be explained by differences in protein expression.

RNA is the intermediate step between genes and proteins; it gets transcribed from DNA molecules, storing the nucleotide sequence from genes in a format that can get translated into proteins. While DNA is identical across all of the cells in an individual's body and rarely changes, RNA levels are highly variable, changing with external factors over time and in different environments. Different human tissues are the result of variable RNA expression across an individual's body; cells in different parts of the body need to perform different functions, and thus express genes in different amounts and will have different RNA levels. For this reason, we have decided to see if RNA expression is an effective predictor of the human organ that cells are from.

2. Dataset description

For this analysis, we decided to use gene expression data produced by the Gentoype-Tissue Expression (GTEx) project. This is a public resource that has worked to collect cell samples from different parts of the human body (54 tissues from 30 different organs) from around 1,000 individuals and measure gene expression levels. These data were produced with a method called RNA-Seq. This method performs sequencing on all RNA transcripts collected from a cell sample, thus helping to identify exactly which genes were expressed in the sample. The number of RNA transcripts with the same sequence in the cell serves as a measure of the expression level of the gene that the RNA was transcribed from. RNA levels are expressed as transcripts per million (TPM), which represents the number of times a particular RNA sequence appeared in the dataset per every million transcript sequenced. The expression level of every gene in the dataset in TPM will serve as the features of our model. In total, the dataset had expression data for 56,200 genes in 17,383 different cell samples.

3. Baseline approach description

In order to make a classifier for the GTEx RNA-Seq dataset, we are using logistic regression. Logistic regression is a supervised machine learning method that, based on certain variables, tries to classify observations into distinct categorical classes. Often those outcome categorical classes (“labels”) are binary, but logistic regression can be modified to also work on multiclass classification. Thus, as the dataset we decided to work with has 30 distinct labels, we sought to use it to evaluate its applicability.

4. Method description

As previously explained, the GTEx RNA-Seq dataset contains measured expression levels for 56,200 different genes across 17,383 samples from 30 human tissues. However, not all genes are differentially expressed across distinct tissues. Thus, our first step to reduce our dataset was to filter out genes with variance less or equal than 10 TPMs. This step guarantees that we would only be working with genes whose expression greatly vary among our samples, meaning that they could be optimal parameters for a classifier.

Next, we further sought to reduce the number of variables of our dataset. To do that, we computed average TPM levels for each gene across all samples, and selected the top 100. This step assures that the predictor variables in our classifier are highly expressed genes. This is important because, as with any data, background noise can be an issue. Consequently, by selecting the top 100 expressed genes, we are confident that the read counts (TPMs) are above background noise.

Lastly, we randomly split our data into training and test datasets (85% and 15%, respectively). With the training dataset, we built and optimized a logistic regression classifier. Applying 5-fold cross validation, we implemented a grid search that evaluated performance of our classifier in order to identify the best values for two hyperparameters: i) using elastic-net, we wanted to identify the best mixing parameter for our model, and thus we tested distinct L1 ratio values ranging from 0.0 to 1.0, with a 0.1 step; ii) additionally, we tested different regularization strength values (C), ranging from 0.1 to 1.0, with a 0.1 step. We used the best hyperparameters in our classifier to evaluate its performance in the test dataset.

5. Evaluation

First, we sought to find the best hyperparameters for our classifier. Using 5-fold cross validation, we found that the best L1 ratio is of 0.5 and C of 0.3. Then, using those values, we applied our classifier to the test data and evaluated its performance. Overall, we had a precision of 0.96, and an average F1-score (across all labels) of 0.83. However, although our test dataset contained 2,695 samples, tissues in our dataset had different sample sizes - for instance, Brain had 426 samples, while Bladder had only 1.

If we take sample size into consideration, the weighted average F1-score of our classifier is of 0.96.

Additionally, we also had results specific to the tissues in the test dataset (Table 1). Among all 30 tissues, 5 of them had a F1-score of 1 (Brain, Blood, Nerves, Pancreas, and Spleen). This means that our classifier correctly predicted the label of all samples of those tissues (760 samples in total). In contrast with that, we had 3 tissues that had a F1-score of 0 (Cervix Uteri, Fallopian Tube, Bladder). For the rest of the data, we had 16 tissues with a F1-score in between [0.90, 0.99] and 6 tissues with F1-score in between [0.68, 0.84].

Table 1 - Tissue-specific performance assessment

Tissue	Precision	Recall	F1-score	# Samples
Brain	1	1	1	426
Pancreas	1	1	1	42
Blood	1	1	1	148
Nerve	1	1	1	104
Spleen	1	1	1	40
Muscle	0.99	0.99	0.99	112
Skin	1	0.99	0.99	273
Lung	1	0.99	0.99	85
Blood Vessel	0.99	0.97	0.98	214
Heart	0.98	0.99	0.98	126
Pituitary	1	0.94	0.97	34
Uterus	0.92	1	0.96	23
Colon	0.95	0.97	0.96	148
Testis	0.92	0.98	0.95	49
Thyroid	0.94	0.95	0.94	107
Ovary	0.94	0.94	0.94	31
Liver	0.9	0.97	0.94	38
Esophagus	0.89	0.98	0.93	207

Stomach	0.98	0.87	0.92	60
Adipose Tissue	0.86	0.96	0.91	184
Adrenal Gland	0.88	0.95	0.91	39
Prostate	0.83	0.85	0.84	34
Salivary Gland	0.96	0.73	0.83	30
Small Intestine	0.92	0.71	0.8	31
Breast	0.88	0.67	0.76	67
Kidney	0.8	0.62	0.7	13
Vagina	0.88	0.56	0.68	25
Cervix Uteri	0	0	0	3
Fallopian Tube	0	0	0	1
Bladder	0	0	0	1

6. Discussion

Overall, our performance assessment results show that we have a fairly accurate model that successfully classified distinct RNA-Seq samples into their tissues of origin. In our tissue-specific analysis, we noticed that our model had better performance with certain tissues than others. This is not unexpected, as many previous works, including from the GTEx Consortium [1], show that tissues form distinct clusters in pairwise clustering analysis based on gene expression. For instance, Brain regions (such as frontal cortex, hypothalamus, amygdala etc.) always form a tight cluster in pairwise analysis, meaning that samples from those regions have similar gene expression profiles. In agreement to that, all 426 Brain samples in our test data were correctly assigned the “Brain” category label. Another tissue that also exemplifies that is Whole Blood, which usually does not cluster with any other tissue in pairwise clustering analysis due to its unique gene expression profile - thus, it was not a surprise that our model correctly classified all 148 Blood samples in the test data. To assess clusters in our test data, we used a dimension reduction technique often used in single-cell RNA expression analysis, UMAP, and colored data points based on their actual label (Appendix figure 1) or predicted label (Appendix figure 2). Our results show the formation of single-tissue clusters, but also clusters with many tissues in them.

However, we also noticed some poorly classified tissues. Namely, we can mention Cervix Uteri, Fallopian Tube, and Bladder. We hypothesize that there may be two reasons why our model did not perform well with those tissues. First, we think that perhaps those tissues have similar gene expression profile with other tissues, and thus samples may have been given incorrect labels. For instance, maybe Cervix Uteri and Fallopian Tube samples were assigned the “Uterus” label - and that could explain why Uterus had a recall of 1, but precision of 0.92. The second hypothesis is that those tissues were poorly sampled in the training dataset, which is a direct consequence of the GTEx sampling bias. For instance, in GTEx there are tissues with many samples in the full dataset, such as Whole Blood (670), Lung (515), and Skeletal Muscle (706). In contrast with that, Cervix Uteri has 19 samples in the whole dataset, Fallopian Tube has 8, and Bladder has 21. Consequently, perhaps the model was not well-trained enough to predict samples belonging to those tissues.

7. Conclusion

In summary, our results show that logistic regression is a successful machine learning method to classify tissues based on gene expression data. Selecting which genes are used as predictors is an important step in making the model, and we showed that prioritizing highly expressed genes, with high variance across samples, is a good feature selection method. Furthermore, we saw that different samples had distinct performances, which is in agreement with the biology behind the data, and the correlation that exists among gene expression profiles between diverse tissues. Moreover, we saw that sampling bias in the training dataset may heavily impact the performance of our model for some tissues, which suggests that one way to boost prediction performance of our model would be to increase sample size for underrepresented tissues.

8. Appendix

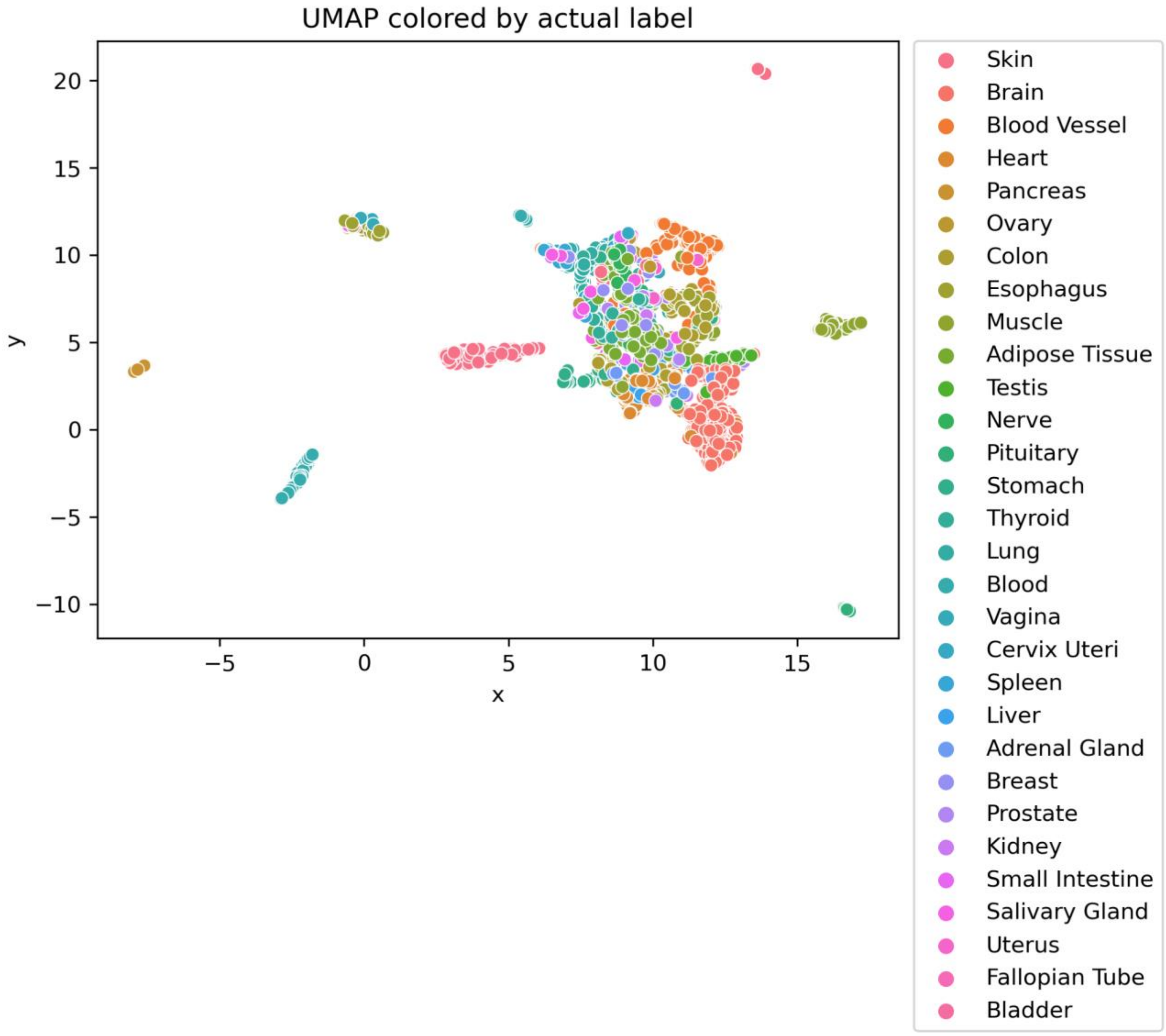

Figure 1. UMAP projection of samples in the test data colored by their actual tissue label.

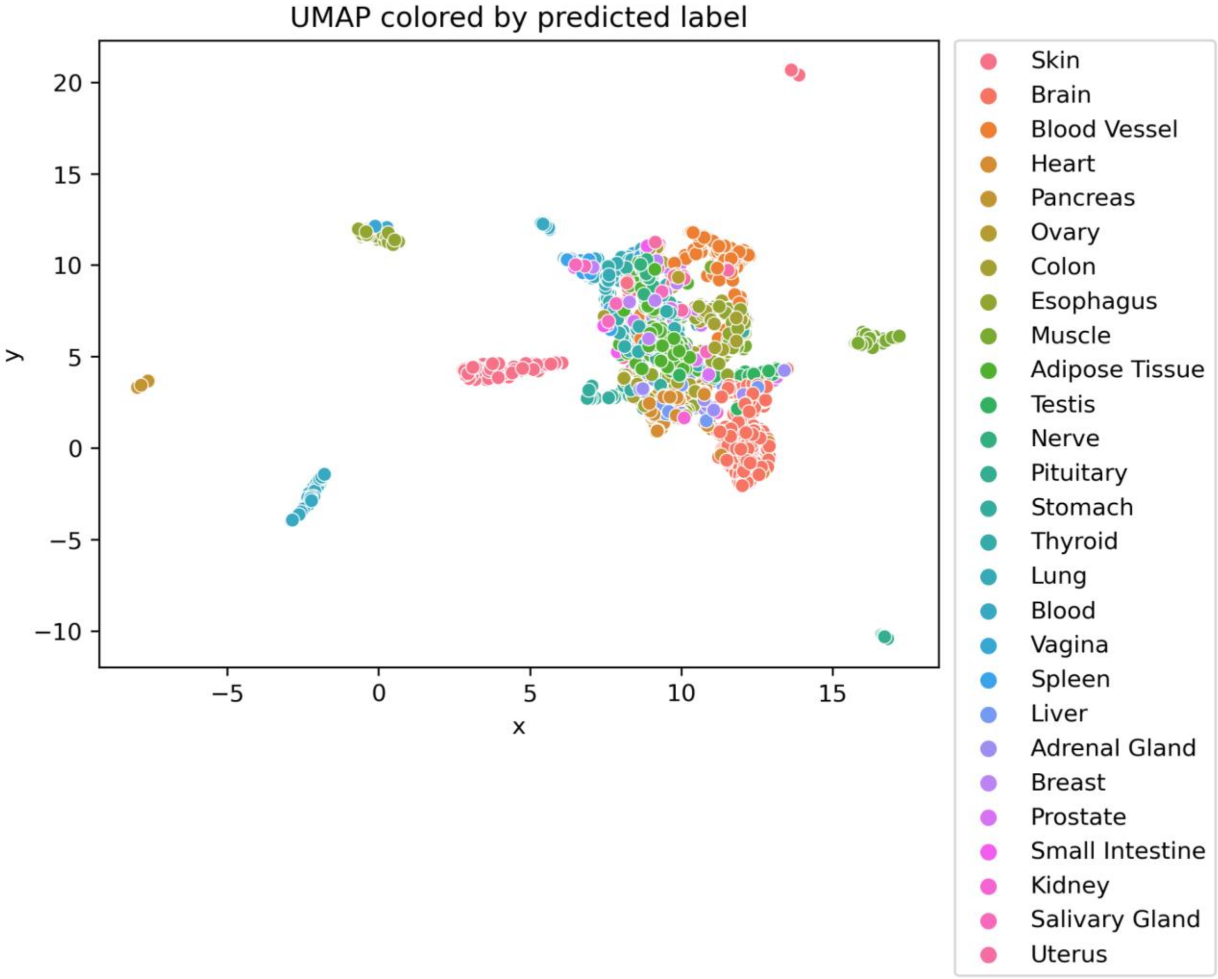

Figure 2. UMAP projection of samples in the test data colored by their predicted tissue label.