An attempt at Leukemia Classification

Andrew Wei
Paul G. Allen School of
Computer Science & Engineering
nowei@cs.washington.edu

Abstract

We find gene expression patterns that exist within specific subtypes of Leukemia and attempt to classify specific subtypes of Leukemia with the intent of analyzing learned patterns from the results.

Introduction

Leukemia is a type of blood cancer that produces excess, abnormal white blood cells that inhibit the production of normal blood cells. There are 17 subtypes of Leukemia and each subtype is characterized by a different set of regulatory genes. Instead of identifying the gene regulatory networks, we analyze the effects of these regulatory networks as we believe they can give us important clues about the impacts and causes of certain subtypes of Leukemia. With this information, we may have a better idea of the genes involved in the regulatory networks of different subtypes of Leukemia, allowing us to better understand these gene regulatory networks and the impactful genes involved in the disease.

Name	Count	Name	Count
MDS	207	Non-Leukemia and healthy bone marrow	73
CLL	448	c-ALL/Pre-B-ALL with t(9;22)	122
AML complex aberrant karyotype	52	AML with t(8;21)	40
AML with normal karyotype + other abnormalities	347	ALL with hyperdiploid karyotype	40
c-ALL/Pre-B-ALL without t(9;22)	237	ALL with t(1;19)	36
T-ALL	174	Pro-B-ALL with t(11q23)/MLL	70
CML	76	AML with t(15;17)	37
AML with t(11q23)/MLL	38	AML with inv(16)/t(16;16)	28
ALL with t(12;21)	58	mature B-ALL with t(8;14)	13

Table 1. Distribution of data

Code and supporting documents: https://github.com/nowei/leukemia-classification
Some extra diagrams and tables are linked externally due to space reasons.

We use array-normalized data from the Microarray Innovations in Leukemia study (MILE) [1]. This dataset contains data from 2096 blood and bone marrow samples from patients and has 17,788 genes with 18 classes, 17-subtypes of Leukemia and 1 group of healthy patients.

There is a data imbalance that we will not directly address in this project, but we note that this may make it harder to learn the important features for classifications of subtypes with a smaller sample size. The small number of samples for some subtypes may cause the model to fit more of the noise than the significant genes involved in the disease. We also note that with the large number of genes, it will be hard to determine the impact of specific genes and verify the results. This problem is challenging because with the rise of machine learning and deep learning, numerous different methods will likely lead to reasonable results, so it is hard to determine which method to use and how to interpret the results. Thus, we choose to use relatively simpler methods to maintain the interpretability of the results.

Background

Cancer classification on gene expression data using machine learning techniques is not new. There are many published papers with new techniques, and applications of old techniques to new datasets that come out every year. We surveyed a handful of papers to get a better idea of different approaches people have tried with gene expression data.

Cascianelli et. al used PCA on the PAM50 dataset to examine whether the subtypes of breast cancer were separable from each other [2]. They found that they could not separate the subtypes by the first two principle components and hypothesized that due to this non-separability, the boundary between subtypes changes depending on the mixed traits. Using Average of Within Class Averages (AWCA), they were able to get a single number that represents a subtype. This method produced relatively decent results with around 90% accuracy. They then went on to try different machine learning approaches and found that regularized multiclass Logistic Regression performed the best.

Danaee et. al used stacked denoising autoencoders to deal with the high dimensions and noisy inputs of gene expression data [3]. The purpose of using an encoder is to denoise the inputs and to try to find the genes that play an important role in breast cancer. They multiply the learned autoencoder weights together to get a rough sense of how the gene expression data interacts, saying that the most heavily weighted ones have the most significant impact and were the most predictive genes. By using the heavily weighted genes as the inputs to a deep learning models, they were able to achieve slightly less accuracy than with the features found using stacked denoising autoencoders but argue that it was more readily interpretable.

Mei et. al trained a stacked autoencoder to provide input to a neural network trained on a subset of the TCGA dataset to identify survival duration for patients with Acute Myeloid Leukemia [4]. They include features like age, cytogenetics, and mutations and achieve results of around 81% accuracy for predicting whether patients survived for more of less than 730 days. They mention that the main limitation to their model was the amount of training data available.

Castillo et. al try to extract differentially expressed genes that help identify different forms of Leukemia [5]. Use the minimum-Redundancy Maximum-Relevance feature selection algorithm, they train SVM, Random Forests, K-Nearest Neighbors, and Naive Bayes classifiers. Afterwards, they extract Differentially Expressed Genes (genes that help differentiate between classes) by testing p-values, log-fold change (measures change in gene expression level), and coverage (whether it helps identify between some number of classes or not). They conclude that

K-Nearest Neighbors with a subset of gene expression data worked the best, followed by SVM and Naive Bayes, with the worst being Random Forests.

Krivtsov et. al outlines procedures on how to extract the gene expression data and utilize some form of either hierarchical or K-means clustering and use permutation analysis to determine significance for the purpose of profiling Leukemia Stem Cells [6]. They use the GenePattern software to distinguish marker genes that would distinguish the two groups, comparing means, using a signal-to-noise statistic, and evaluating significance. They extracted and recorded the data themselves instead of taking the data from a database.

These papers showed that there were many ways to approach the problem of cancer classification using gene expression data and have inspired us to take a somewhat similar approaches, with an emphasis on finding DEGs or significant genes/features.

Results and Methods

Methods

We take a multifaceted approach to this problem of identifying significant genes and classifying Leukemia subtypes. We use a 90/10 train/test split to leave some data for evaluation purposes.

To look for significant features within each Leukemia subtype, we perform Welch's t-test using the training data and compare each subtype against the *Non-Leukemia and healthy bone marrow* patients. Then we apply the Bonferroni correction with $p = 0.05$ and $m = 17{,}788$ to correct for multiple hypotheses. Afterwards, we determine the similarity of significant features between Leukemia subtypes to get a better idea of how similar the selected genes were between Leukemia subtypes, so we looked at the Jaccard similarity (a set similarity metric) between the selected significant features between Leukemia subtypes. We then look up some of the top shared features to confirm the pattern was also observed by others.

Afterwards, we attempt to classify the Leukemia subtypes by using multi-class Logistic Regression, which uses a 1 vs. all classification scheme and has a model with learned weights for each class. This makes it easier for the model to classify the classes with a smaller sample size, since the model will be able to learn what is not a specific class by learning from the remaining data. We focus on using L1 regularization because we believe that not all gene expressions are relevant to determining whether someone has a specific subtype of Leukemia or not. We varied our training by using all of the features and just using the combined significant features we found, along with three normalization schemes: not normalizing, normalizing by the training data, and normalizing by the healthy patients in the training data. The same normalization schemes are applied when evaluating the model. We use 5-fold cross-validation to get a sense of the generalizability of our different models. We then look at the similarity between the learned weights of the learned models as well as the learned weights themselves to examine the most impactful genes found by the model. We conclude by comparing the classification accuracy of logistic regression with K-Nearest Neighbors. We vary the number of neighbors and whether we use all features or just the significant features.

We believe that our approach will work well on this problem because it allows us to tackle the problem of identifying the significant genes from multiple directions to arrive at a

better understanding of which genes are significant indicators of each subtype of Leukemia. Our initial methods differ from existing methods in that we perform our analysis on 17 subtypes of Leukemia, while previous research we have seen only does this analysis among 4 subtypes of Leukemia [5] and do not focus on the similarities of the significant features. Our latter methods do not differ much from existing methods in that it focuses on using a machine learning algorithm as a means of classification.

We evaluate our ability to locate significant features by checking some of the genes against existing research to confirm that these patterns were also observed by researchers. This leaves us with many genes that could be involved in the gene regulatory networks that may or may not have been observed by other researchers that could be a focus within future research on gene regulatory networks for specific subtypes of Leukemia. We evaluate our classification results on the top-1 and top-5 accuracy because there are 18 total classes, so top-5 accuracy gives us a sense of how close our model was to predicting the correct class.

Results

After selecting out the significant features for each subtype of Leukemia, we found a total of 1,408 significant features. We also examined the Jaccard similarity of the found significant features and learned that *MDS* seemed to be the most different subtype of Leukemia. This may require some more investigation in the future. The number of labels that share a particular significant feature range from 1 to 16, so we can look at a handful of the top shared and least shared labels for significant genes and look for them in existing research.

Gene	Description	# of shared labels	Observed in
10487_at	CAP1	16	[7]
1116_at	CHI3L1	15	[8]
116362_at	RBP7	15	[9]
10123_at	ARL4C	15	[10]
1118_at	CHIT1	15	[11]
114880_at	OSBPL6	1 (MDS)	[12]
100128907_at	hypothetical protein LOC100128907	1 (MDS)	Not found
10006_at	ABI1	1 (CLL)	[13]
100128309_at	hypothetical protein LOC100128309	1 (CLL)	Not found
100129015_at	hypothetical protein LOC100129015	1 (CLL)	Not found

Table 2. Selected results from the search for significant features

We observe that some hypothetical proteins were marked as significant genes, so further studies surrounding these genes could be carried out. There are likely other genes that have not been observed in studies and studying these genes might give us a better idea about the regulatory networks of different Leukemia subtypes.

Then we tried multi-class Logistic Regression. Our 5-fold cross-validation results showed that not normalizing had the best top-5 accuracy, while normalizing by healthy patient data in the training dataset had the best top-1 accuracy. We decided to train a model with all three normalization schemes because training the models take a relatively short amount of time. When training the models on the full data, we notice that each model is able to learn the training data completely. We present the test results below:

Normalization Scheme	**w/ all features**		**w/ significant features**	
	Top 1 acc.	**Top 5 acc.**	**Top 1 acc.**	**Top 5 acc.**
Don't normalize	**0.919**	**1.0**	**0.881**	**0.990**
Normalize across entire training dataset	0.919	0.995	0.857	0.981
Normalize by healthy patient data in training dataset	0.900	0.995	0.843	0.971

Table 3. Accuracy of multi-class Logistic Regression models on test data

It is reasonable that the accuracy would be higher with all the features, since significant features don't include features that indicate that it isn't some class, which forces the model to find indications that it isn't some class from the remaining significant features for other classes. We also note that not normalizing seems to perform the best, which could be caused by some genes being strictly related to the inhibition or presence of certain Leukemia subtypes, so not normalizing further actually better expresses the degree in which some subtypes are present or not. Since we used L1 regularization, we have seen that many of the features in the models are zero-weighted, which seems to enforce our assumption that only a minority subset of gene expression data will actually be relevant to classifying between Leukemia sub-types. We notice that on the confusion matrix with all the features and no normalization, the errors we make are pretty sparse with the only readily visible pattern being that we misclassify *c-ALL/Pre-B-ALL without t(9;22)* relatively frequently. These errors might give us some clue as to the feature space that these classifications live in, e.g. the misclassifications could be due to how close *c-ALL/Pre-B-ALL without t(9;22)* is to *ALL with hyperdiploid karyotype* within the feature space.

We can also check the heavily weighted weights our models found. These weights can give us a clue as to which features were most significant in determining whether a patient had a specific subtype of Leukemia or not. To do this, we would have to look at the data and see whether patients had generally positive or negative records for the heavily weighted features within each subtype's classification model. If the product is generally positive for a subtype, then we know that it is a strong indicator of the specific subtype and if it is generally negative, we know that the feature was a strong indicator that it was not that subtype. As an example, we look at the most heavily weighted weights of *mature B-ALL with t(8;14)*. We choose *mature B-ALL with t(8;14)* because this class only has a sample size of 13, making it easier to analyze the patterns and look for general trends.

Weight	Name	Description	Values generally positive/negative	Product sign	Strong indicator
0.275562	151126_at	ZNF385B - zinc finger protein 385B	Positive	Positive	For
0.204056	284013_at	VMO1 - vitelline membrane outer layer 1 homolog (chicken)	Positive	Positive	For
-0.11895	7006_at	TEC - tec protein tyrosine kinase	Positive	Negative	Against
-0.12564	100131601_at	similar to hCG1980470	Positive	Negative	Against

Table 4. Heavily weighted values of learned weights and whether they are strong indicators for or against *mature B-ALL with t(8;14)* classification.

We also looked at the cosine similarity between learned weights for the best-performing model. We do this to get an idea of how similar the learned weight vectors are. We observe that the majority of the learned weights are generally independent, with *Non-leukemia and healthy bone marrow* being the most opposed to *MDS* and *c-ALL/Pre-B-ALL with t(9;22)* being the most opposed to *c-ALL/Pre-B-ALL without t(9;22)*. Even though the learned weights were mostly independent, this data could give us some context for the errors that our model made. For example, most of the misclassifications our model made are between classes with relatively more opposite weights, which seems to indicate that the model cannot correctly distinguish between these classes using the learned weights due to their sharing of weighted features. Since we learn the training set completely, this error may also be attributed to the variance between the training and test sets, meaning that our learned model fails to generalize. While this model performs decently well, with above 90% accuracy, we believe that it could not learn the underlying patterns under the data well enough to generalize to unseen data, but it is a good model for learning about specific features to pay attention to for when embarking on future endeavors.

Then when trying K-Nearest Neighbors as a point of comparison, we notice that our results were not as impressive as the ones we obtained through multi-class Logistic Regression.

# of neighbors	w/ all features		w/ significant features	
	Top 1 acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
3	0.767	0.881	0.790	0.895
5	0.771	0.914	0.790	0.929
10	**0.795**	0.962	**0.819**	**0.962**
15	0.786	0.967	**0.819**	0.952
20	0.790	**0.971**	0.786	0.952

Table 5. Performance of K-Nearest Neighbors.

We observe the best top-5 performance when using all the features and with 20 neighbors and the best top-1 performance using just the significant features with either 10 or 15 neighbors. Looking at the errors in the confusion matrix for 10 neighbors with only significant features, we see some of the same errors that the multi-class Logistic Regression model made, along with new errors. This is likely due to the complex feature space in which these classifications live, along with the low number of samples and the high number of dimensions, which induces sparsity. This means that we will suffer from many errors due to how spread out the data is, especially

when classifying patients with data that lie near the boundaries between classes. The results seen here go against what was observed by Castillo et. al in [5], but this may be due to their more general classification scheme for subtypes of Leukemia (i.e. having 4 subtypes compared to the 17 we use).

Conclusion

We address the computational challenge of having a data imbalance by using a 1 vs. all multi-class Logistic Regression scheme, which allowed our model to learn the signs that a sample is not a particular class from the other data. We also address the problem of finding significant features among all the different classes by focusing on the significant features on a per-class basis, which allowed us to break the problem down into more computationally more tractable steps.

In the future, we plan on taking a deeper dive into the significant features and the heavily weighted features we found, which may give us more insight into genes involved in the regulatory networks of different subtypes of Leukemia. We may also try an alternative 1 vs. all classification scheme where we create a separate model for each Leukemia subtype and train them on the significant features we found. Trying different/tuning regularizers and utilizing other machine learning or deep learning methods may help us overcome the problems we saw with our Logistic Regression model. We may also attempt to map out the gene regulatory networks of different subtypes of Leukemia using probabilistic graphical methods. As another way to address the data imbalance, we may be able to reattempt these experiments when more Leukemia patient data has been collected and published.

By performing simple statistics on gene expression data and observing significant deviations, we can learn enough information to start seeing how gene expressions differ between healthy and sick patients. This may give us clues when attempting to map out the gene regulatory networks involved in different subtypes of Leukemia. Additionally, we have seen that multi-class Logistic Regression produces relatively decent results when tasked with identifying different subtypes of Leukemia in the MILE dataset. We have also seen that K-Nearest Neighbors suffers in this setting, likely due to the sparsity of the available data. We conclude by noting that with new models being made and more data being collected every day, we may become more inclined to try increasingly complicated approaches to solve problems. This may lead to making relatively easy problems harder than they actually are. Thus, it is important to begin approaching problems with simple, interpretable methods, expanding on those methods, and analyzing their results before diving into more complex methods.

Bibliography

[1] Kohlmann, Alexander et al. “An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in Leukemia study prephase.” British journal of haematology vol. 142,5 (2008): 802-7. doi:10.1111/j.1365-2141.2008.07261.x

[2] Cascianelli, S., Molineris, I., Isella, C. *et al.* Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Sci Rep* **10,** 14071 (2020). https://doi.org/10.1038/s41598-020-70832-2

[3] Danaee, P., Ghaeini, R., Hendrix, D. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. PSB (2017).

[4] Mei Lin, Vanya Jaitly, Iris Wang, Zhihong Hu, Lei Chen, Md. Amer Wahed, Zeyad Kanaan, Adan Rios, & Andy N. D. Nguyen. (2018). Application of Deep Learning on Predicting Prognosis of Acute Myeloid Leukemia with Cytogenetics, Age, and Mutations.

[5] Castillo, Daniel et al. "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level." *PloS one* vol. 14,2 e0212127. 12 Feb. 2019, doi:10.1371/journal.pone.0212127

[6] Krivtsov, Andrei V et al. "Gene expression profiling of Leukemia stem cells." *Methods in molecular biology (Clifton, N.J.)* vol. 538 (2009): 231-46. doi:10.1007/978-1-59745-418-6_11

[7] Xie, Shuanshuan et al. "Systematic analysis of gene expression alterations and clinical outcomes of adenylate cyclase-associated protein in cancer." Oncotarget vol. 8,16 (2017): 27216-27239. doi:10.18632/oncotarget.16111

[8] Bergmann OJ, et al. High serum concentration of YKL-40 is associated with short survival in patients with acute myeloid leukemia. Clin Cancer Res. 2005;11(24 Pt 1):8644–8652.

[9] Qu, Ying, et al. "Novel Gene Signature Reveals Prognostic Model in Acute Myeloid Leukemia." Frontiers in Genetics, vol. 11, Oct. 2020, p. 566024. DOI.org (Crossref), doi:10.3389/fgene.2020.566024.

[10] Engel, T. et al. https://www.wikigenes.org/e/gene/e/10123.html?vs=1

[11] Cho, Soo Jung et al. "Chitotriosidase in the Pathogenesis of Inflammation, Interstitial Lung Diseases and COPD." Allergy, asthma & immunology research vol. 7,1 (2015): 14-21. doi:10.4168/aair.2015.7.1.14

[12] Figueroa, Maria E et al. "Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features." Blood vol. 113,12 (2009): 2795-804. doi:10.1182/blood-2008-08-172387

[13] Taki T, Shibuya N, Taniwaki M, Hanada R, Morishita K, Bessho F, Yanagisawa M, Hayashi Y. ABI-1, a human homolog to mouse Abl-interactor 1, fuses the MLL gene in acute myeloid leukemia with t(10;11)(p11.2;q23). Blood. 1998 Aug 15;92(4):1125-30. PMID: 9694699.