Partial Least Squares and Logistic Regression Analysis of MicroRNA Expression Profiles for Breast Cancer Detection

By: Jessica Murphy

INTRODUCTION

MicroRNAs (miRNAs) are small, noncoding RNAs that regulate gene expression. Their alteration has been associated with several types of human cancer and their expression profiles have been used to classify tumors[1].

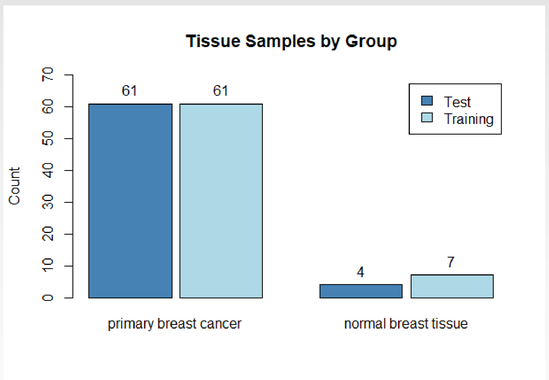
In a recent paper by Matamala et al., miRNAs deregulated in breast tumors were identified using differential expression analysis and support vector machines (SVM). These miRNAs were compared with circulating miRNAs in plasma to determine their potential as noninvasive biomarkers for early breast cancer detection[1].

OBJECTIVES

1. Predict whether a breast tissue sample is cancerous or normal based on miRNA expression profiles.
2. Determine which miRNAs are most representative in this predictive model.
3. Compare these results to those found in the paper.

DATA

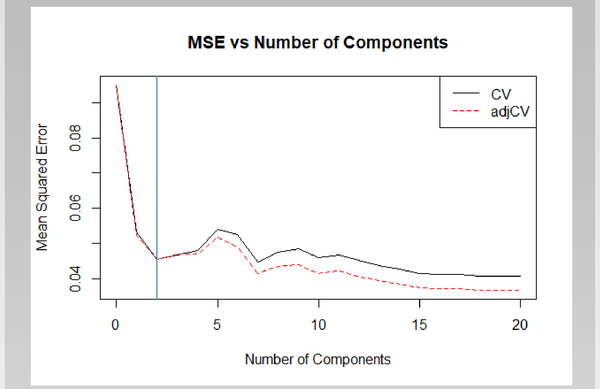
The dataset from the paper contains the expression profiles of 1,919 miRNAs in breast tissue samples from 133 different women. These samples were divided into training and test sets according to the plot below.



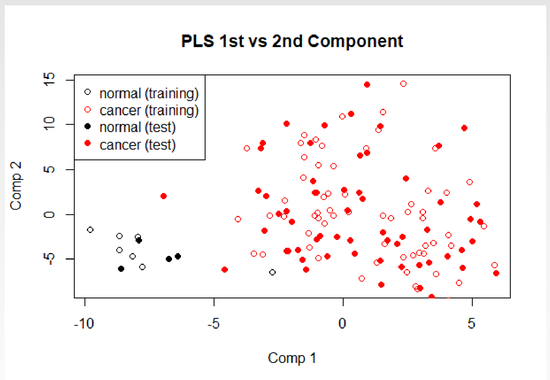
The data was then filtered to remove miRNAs with low expression variation across samples (< 0.03), resulting in 699 remaining miRNAs. The predictors in both the training and test sets were centered according to the predictor means of the training set.

METHODS

Partial least squares (PLS) was performed on the training data and cross-validation was used to determine the optimal number of components to include[2]. The blue line on the following plot shows that two components were chosen.



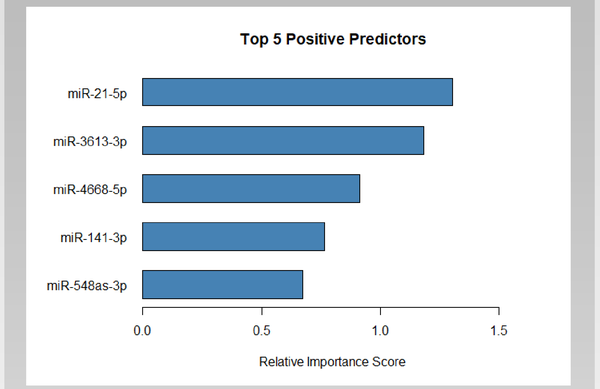
The test data was then projected as additional samples onto the selected components from the training data[3]. A plot of the two components for both the training and test sets can be seen below, colored by tissue sample. The plot shows the types of tissue are fairly well separated by these two components, with slightly more separation seen in the training set than the test set.

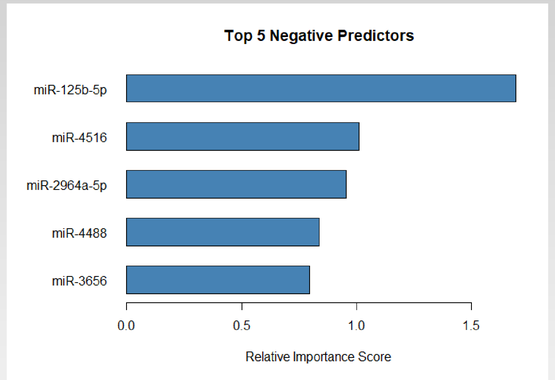


A logistic regression model was also built on the training set using these two PLS components. The accuracy, specificity, and sensitivity of the model were evaluated on the test set and compared to the performance of the paper's SVM classifier.

RESULTS

The relative importance of each miRNA in the PLS model was calculated by normalizing the model coefficients[4]. These normalized scores were ordered and the top positive and negative predictors from the model are shown in the plots below.





All of these top predictors, except for miR-2964a-5p, were cited in the paper as having the greatest difference in expression between breast tumors and healthy breast tissues. Four miRNAs (miR-125-5b, miR-3613-3p, miR-4668-5p, and miR-3656) were also included in the signature generated by the paper's SVM classifier[1].

Both miR-125b-5p and miR-21-5p have been repeatedly associated with breast cancer, but six miRNAs (miR-3613-3p, miR-4668-5p, miR-4516, miR-548as-3p, miR-4488, and miR-3656) had not been associated with breast cancer before. Three (miR-21-5p, miR-125b-5p, and miR-3656) were also found to be differentially expressed in plasma[1].

RESULTS Cont.

The logistic regression model built using the PLS components predicted perfectly on the training set and only misclassified 1 sample out of 65 from the test set. These results are compared with the paper's SVM classifier in the table below.

	PLS		SVM	
	Training	Test	Training	Test
Accuracy %	100	98	98	100
Sensitivity %	100	98	100	100
Specificity %	100	100	83	100

The SVM classifier surprisingly predicted perfectly on the test data rather than the training data.

CONCLUSIONS

- The performance of the PLS classifier is very comparable to that of the SVM classifier.
- PLS with just two components produced an almost identical list of top miRNAs as the paper's differential expression analysis.
- PLS produced an additional miRNA (miR-2964a-5p) not selected by the paper that would be worth further analyzing.
- PLS confirmed the role of miR-21-5p and miR-125b-5p in discriminating against breast cancer in this dataset.

REFERENCES

1. Matamala, N., Vargas, M. T., Gonzalez-Campora, R., Minambres, R., Arias, J. I., Menendez, P., … Benitez, J. (2015). Tumor MicroRNA Expression Profiling Identifies Circulating MicroRNAs for Early Breast Cancer Detection. *Clinical Chemistry*, *61*(8), 1098–1106. doi: 10.1373/clinchem.2015.238691
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. New York: Springer.
3. Garcia, S. (2016, December 17). Partial Least Squares for Leukemia Dataset. Retrieved November 26, 2019, from https://rpubs.com/Saulabrm/PLS1.
4. Hoare, J. (n.d.). Using Partial Least Squares to Conduct Relative Importance Analysis in R. Retrieved November 26, 2019, from https://www.displayr.com/using-partial-least-squares-to-conduct-relative-importance-analysis-in-r/.