Statistical Analysis of Glycoprotein Data in Breast Cancer Cell Lines

Spencer Bowen* Alexandra Piryatinska Leslie Timpe

Abstract

Glycoproteins found in the cell membrane have previously been shown to be useful as biomarkers in the diagnosis and treatment of cancer. Our goal is to identify glycoproteins which may serve as biomarkers for several subtypes of breast cancer.

Through mass spectrometry, a large number of glycoproteins in multiple cancer and non-cancer cell lines were measured. The candidates for biomarkers were obtained by variable selection through the use of methods including LASSO Logistic Regression, Random Forest.

Results will be shown for the following breast cancer cell classifications: benign or malignant, basal or luminal, and claudin-low; additionally, the mechanisms used for prediction and variable selection will be explained.

Key Words: LASSO Logistic Regression, Random Forest

1. Introduction

Breast cancer is the most common type of cancer found in American women. Treatment of the disease is difficult due to the disease's heterogeneous nature with regard to type of tumor, chance of recurrence, and expected response to therapy. Therefore it is an important research objective to identify protein biomarkers that provide clinically useful information regarding the diagnosis, prognosis, or response to treatment of breast cancer. It is with this goal in mind that the laboratory of Dr. Bruce Macher of the San Francisco State University Department of Chemistry and Biochemistry have worked to determine if glycoproteins found in cells could be used to diagnosis the disease and distinguish between various subtypes of breast cancer.[6]

Biomarkers such as glycoproteins found in the blood may be able to provide information about a tumor before surgery, they may also used to monitor for recurrence. Glycoproteins are currently being used as biomarkers for prostate cancer; however, no blood test for breast cancer detection is currently in clinical use. If blood tests for diagnosis, prognosis, and monitoring of breast cancer were available that were superior to existing tests, they would be widely used.[6]

The goal of this paper is to use results from LASSO Logistic Regression and Random Forest to identify glycoproteins which may help to diagnosis breast cancer or to distinguish between various subtypes of the disease. The paper will be structured as follows. In Section 2.1, we present our data. In Section 2.2, we present methodology for data analysis. In Section 3 we present our results. In Section 4 we give our conclusions and discuss our results.

*San Francisco State University, 1600 Holloway, San Francisco, CA 94132

2. Methodology

2.1 Data

First Dataset

Twenty-six cell lines were grown in culture dishes by the Macher laboratory. These included twenty-one breast cancer cell lines and five benign tumor cell lines. The Macher's laboratory then used electrospray ionization/tandem mass spectrometry to determine the protein components for each cell line with a total of 431 distinct glycoproteins being observed. These measurements were widely spread, so a log_2 transform was applied to the data.

In addition to the benign and malignant classification, the cell lines were classified as being benign or malignant and whether or not they are claudin-low. These are subtypes which give information about the hormone receptor status of the cell, which is valuable information for treatment of the tumor.

Second Dataset

In the second phase of the analysis, after our preliminary results, the Macher's Laboratory measured proteins (not from the cell membrane) through methods other than mass spectrometry. Four proteins were measured from seven breast cancer cell lines and four benign tumor cell lines; additionally, eight cell lines were classified as basal-like, and five were classified as luminal-like.

2.2 Methodology

In this paper we analyze data such that the number of predictors (431) is sufficiently larger then the number of observations (26). Our goal is to select variables to predict the type of the breast caner. Classical statistical approaches are not applicable to such problem. Therefore we are going to use LASSO Logistic Regression and Random Forest for the variable selection.

LASSO Logistic Regression

LASSO Logistic Regression (LLR) is penalized logistic regression with an L_1 penalty term, see [3]. In this case the following penalized version of log-likelihood function is maximized

$$l(\beta,\lambda) = -\sum_{i=1}^{n}[(1-y_i)\beta^T\mathbf{x}_i' + \log(1+exp(-\beta^T\mathbf{x}_i'))] - \lambda\sum_{i=1}^{m}|\beta_i|.$$

To fit a LASSO Logistic Regression, the R-software package *glmnet*[2] was used. To find the optimal λ the package performs leave-one-out cross-validation. Our best predictors were chosen from the full model with the corresponding optimal λ. To test the accuracy of LLR, we performed leave-3-out cross validation for all possible combinations to produce a model for each training set. LLR gives us a probability, so for these models we computed the sensitivity and specificity by taking the true positives and true negatives their predictors on the respective test sets for a given cut-off value. Then we obtained our best cut-off value

for each classification by selecting the cut-off value which minimized $(1-\text{Specificity})^2 + (1-\text{Sensitivy})^2$.

RandomForest

Random forest is the learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees [1]. We run Random Forest using Package *randomForest*[4] to classify data into two categories. To test the accuracy of Random Forest, we performed leave-3-out cross validation and used them to produce a model for each test set. Sensitivity and Specificity was computed by taking the true positives and true negatives predictors generated by these models on their respective test sets.

Gini criteria was used to find importance of the variables. We then totaled the variable importance over all training models, standardized them and selected the best predictors.

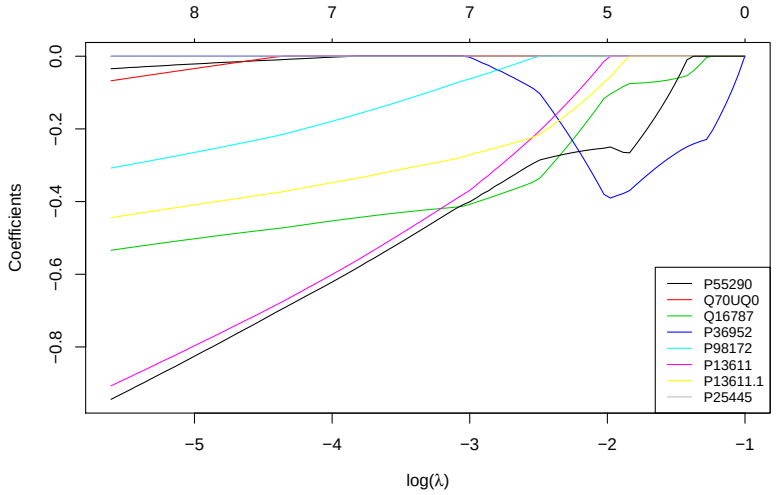
3. Results

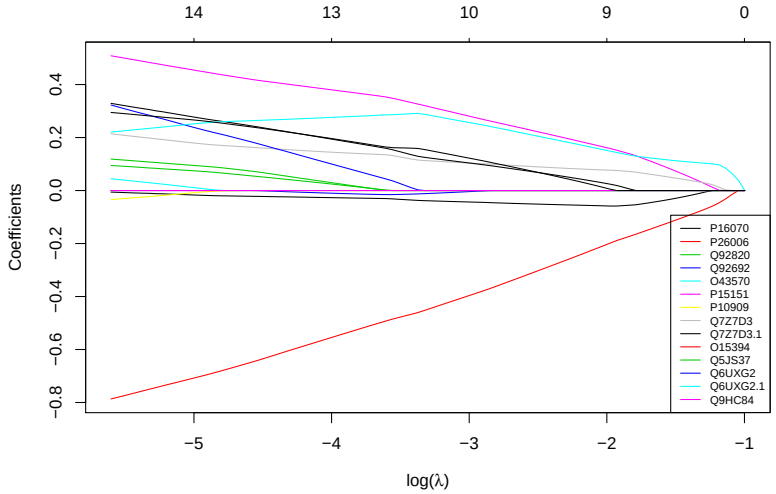
3.1 LASSO Logistic Regression

First *glmnet* was used to obtain cross-validated optimal values for λ for each classification.

Figure 1 shows the coefficient paths with respect to λ, and Figure 2 shows the binomial deviance from cross validation, with the red line representing the optimal value for λ i.e. the value which minimizes the deviance. This is the point where our active predictors from Table 1 are taken.

Then models were created for each classification, with active coefficients shown in Table 1.





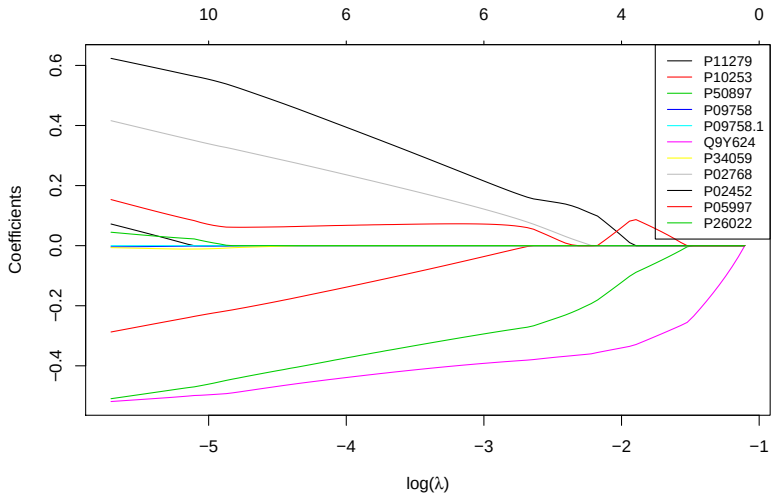


Figure 1: Top to bottom: benign or malignant, basal or luminal, claudin-low. The numbers on top of each graph denote the number of nonzero coefficients. The key shows the glycoprotein corresponding to each coefficient path.

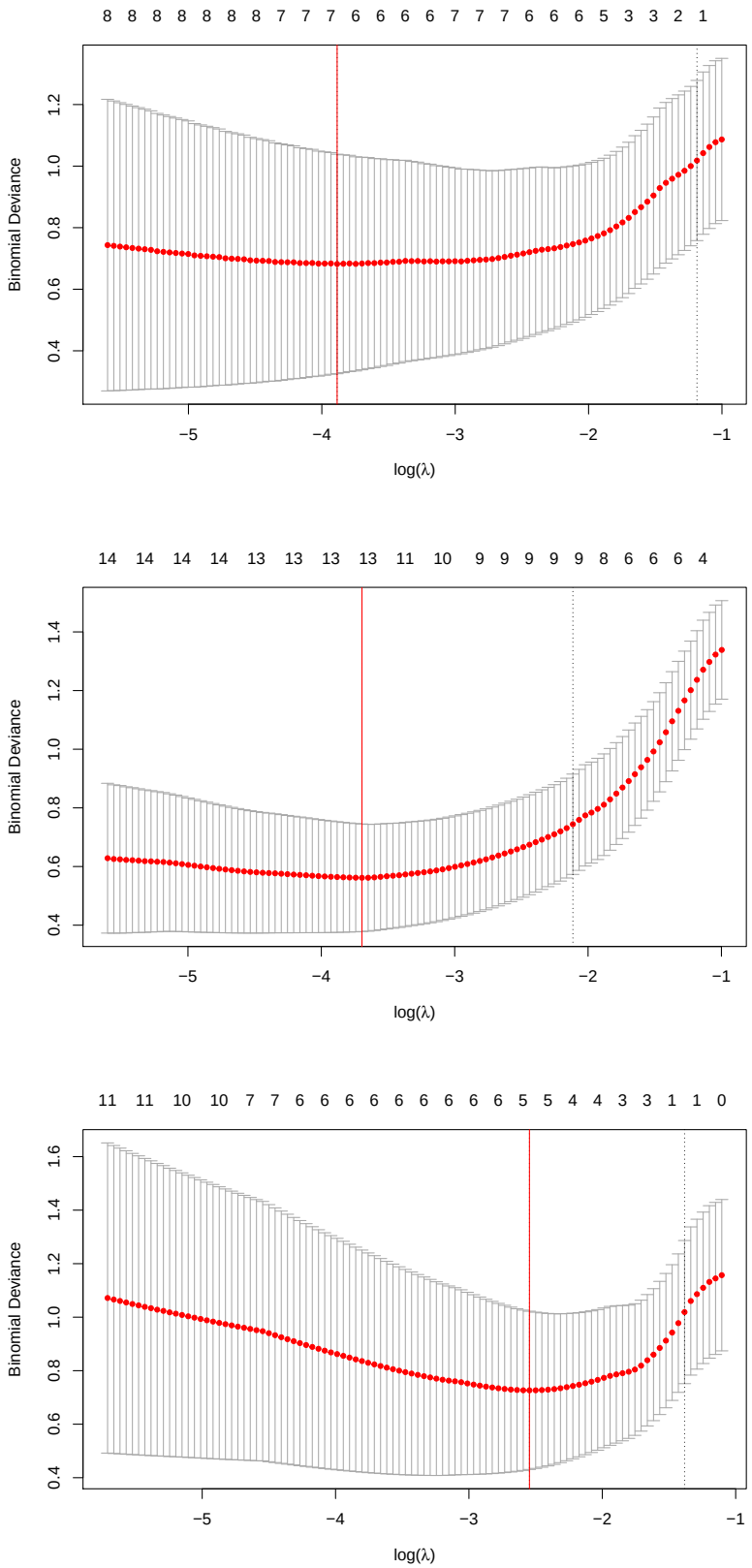


Figure 2: Top to bottom: benign or malignant, basal or luminal, claudin-low. The numbers on top of each graph denote the number of nonzero coefficients. The red line denotes the λ which results in minimized validated deviance.

benign vs. malignant	$\lambda = 0.0312$
Q16787	-0.4287
P36952	-0.1193
P98172	-0.4806
P13611	-0.3090
P13611.1	-0.0000
P25445	-0.5042

basal vs. luminal	$\lambda = 0.0226$
P16070	-0.0287
P26006	-0.5210
Q92820	0.0124
Q9UBG0	-0.0133
O43570	0.3660
P10909	0.1393
Q7Z7D3	0.1756
Q7Z7D3.1	0.0000
O15394	0.0140
Q5JS37	0.0699
Q6UXG2	0.2810
Q6UXG2.1	0.0000
Q9HC84	0.1792

claudin-low	$\lambda = 0.0748$
P50897	-0.2597
Q9Y624	-0.3768
P02768	0.0635
P02452	0.1524
P05997	0.0456

Table 1: Active predictors for each classification at their respective optimal values λ.

The LASSO logistic regression provides us with preditions of the probability of success. But our problem is to classify data. Therefore the optimal cut-off point should be found. For this purpose we use ROC curves. Each point on the ROC Curve depicts the sensitivity and specificity for a given cut-off value.

The red point indicates the cut-off value which gives the best sensitivity and specificity. Sensitivity and specificity for the optimal cut-off value are given in Table 2.

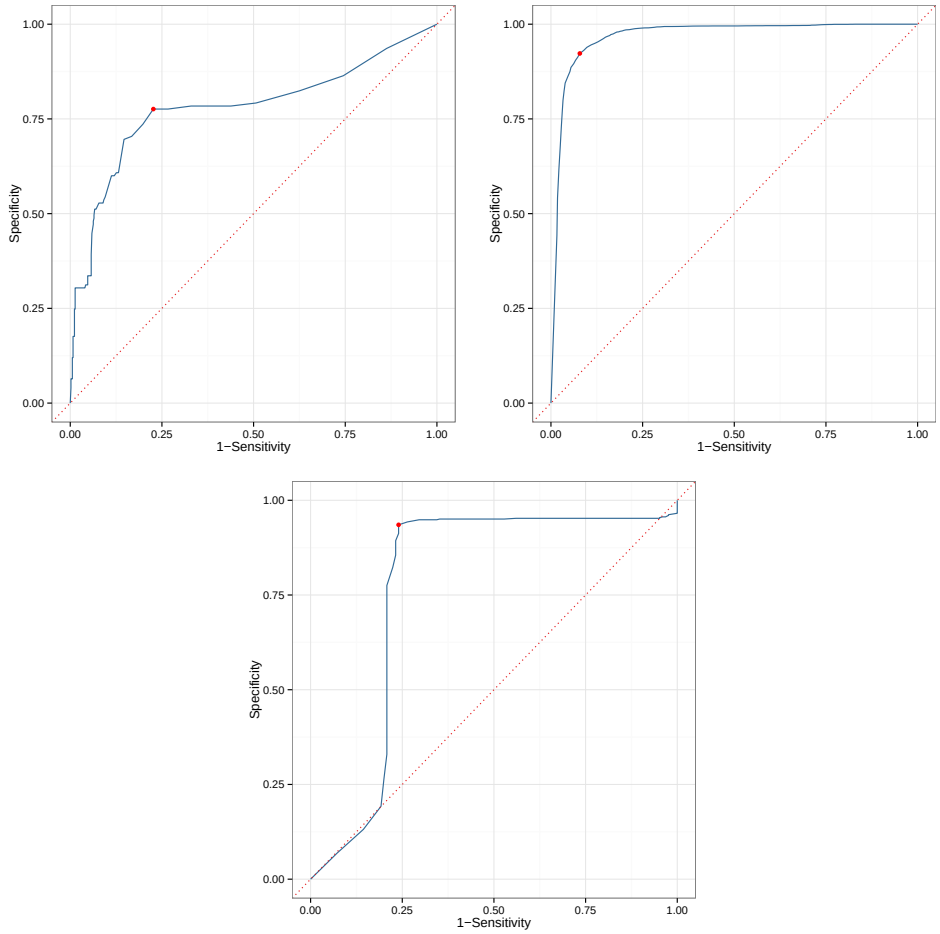


Figure 3: Clockwise: benign or malignant, basal or luminal, claudin-low. The red point marks best cut-off value based on Euclidean distance from (0,1).

	Cut.Off	Sensitivity	Specificity	Accuracy
benign vs. malignant	0.91	0.77	0.78	0.77
basal vs. luminal	0.17	0.92	0.92	0.92
claudin-low	0.19	0.76	0.94	0.90

Table 2: Optimal cut-off value for each classification. Note that benign vs. malignant and claudin-low use leave-2-out cross validation due to few positive cases in the dataset.

3.2 Random Forest

Another approach to select predictors for different types of breast cancer is to use Random Forest. Random Forest provides us with classification while also identifying the important predictors. Figures 4 and 5 show the 15 most important predictors in the ascending order

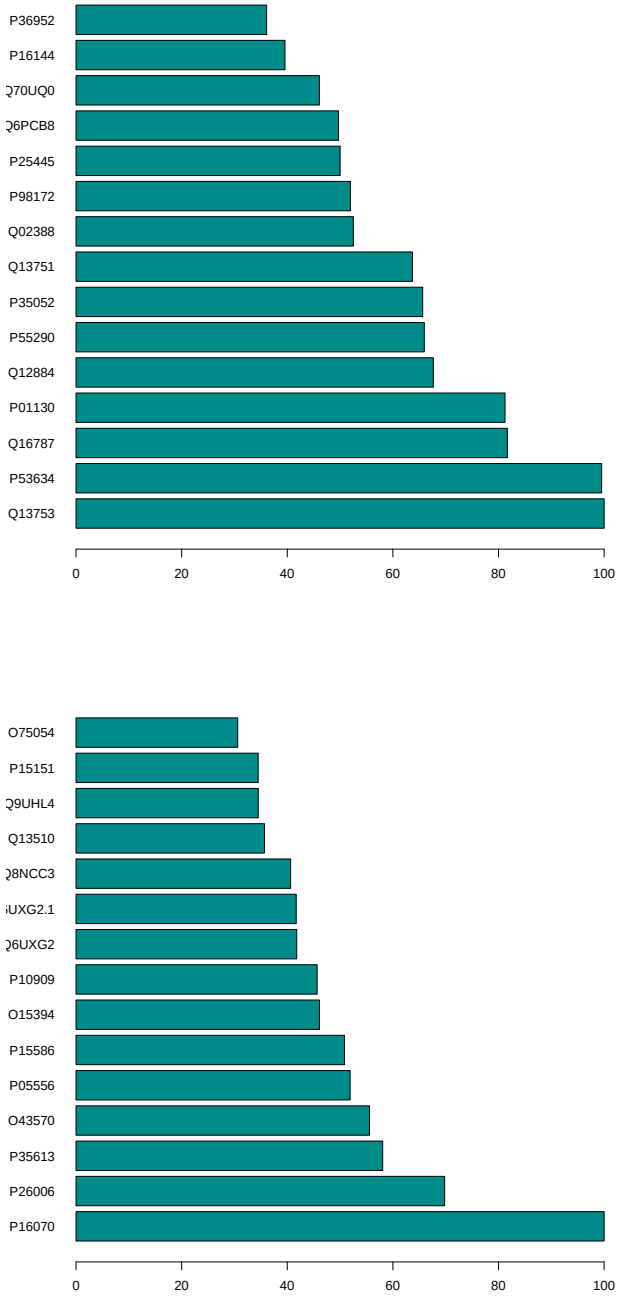


Figure 4: Variable importance for benign or malignant (top) and basl or luminal (bottom) classifications.

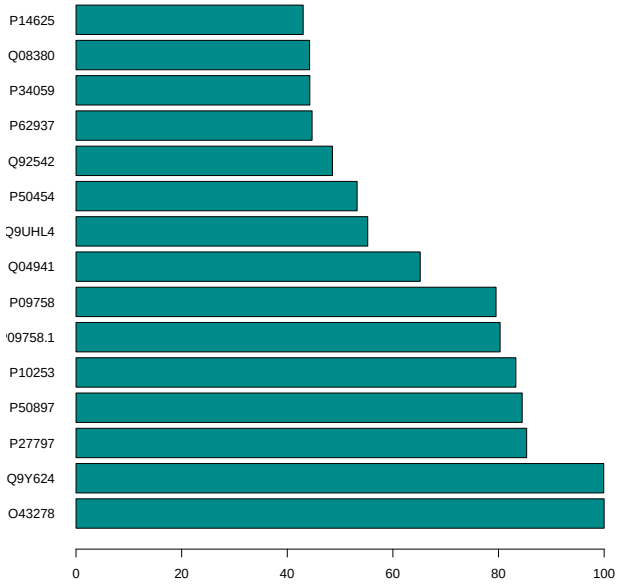


Figure 5: Variable importance for claudin-low classification.

The cross-validated sensitivity and specificity for each Random Forest classification are shown in Table 3.

	Sensitivity	Specificity	Accuracy
benign vs. malignant	1.00	0.49	0.90
basal vs. luminal	0.66	1.00	0.89
claudin-low	0.18	0.99	0.84

Table 3: Results for each classificiation using leave-3-out cross validation.

3.3 Second Dataset

The methodology for above was then repeated for the second dataset. Figure 6 shows the coefficient paths and cross validation from LLR, while Figure 7 shows the specificity and sensitivity for our optimal cut-off value (the red point).

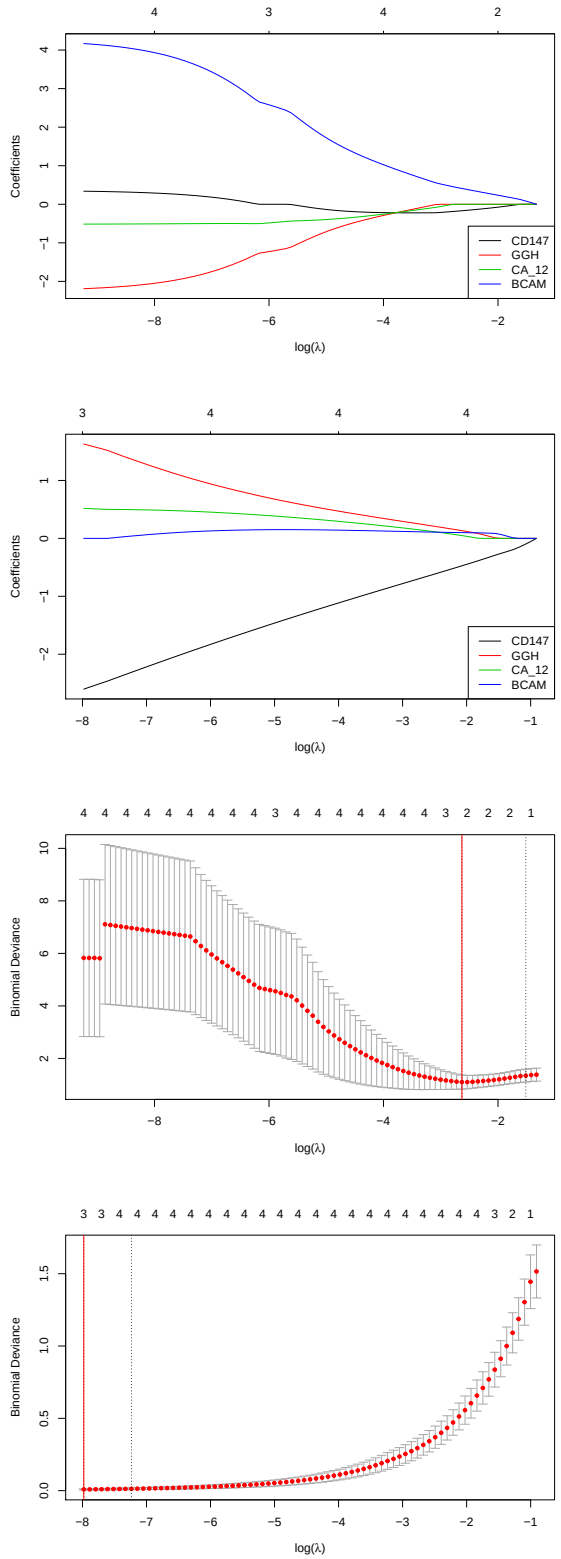


Figure 6: Coefficient paths and optimal λ from cross-validation for benign or malignant (top) and basal or luminal (bottom).

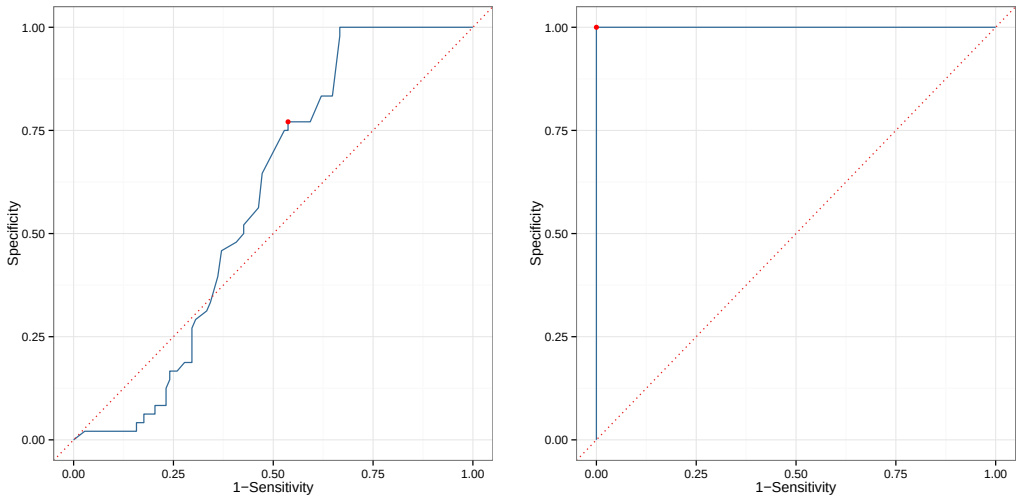


Figure 7: Left: benign or malignant; right: basal or luminal. The red point marks best cut-off value based on Euclidean distance from (0,1).

	Cut.Off	Sensitivity	Specificity	Accuracy
benign vs. malignant	0.75	0.46	0.77	0.56
basal vs. luminal	0.05	1.00	1.00	1.00

Table 4: Results from LLR using leave-2-out-cross-validation

Figure 8 shows our ranking of the important predictors from Random Forest with the table showing our sensitivity and specificity.

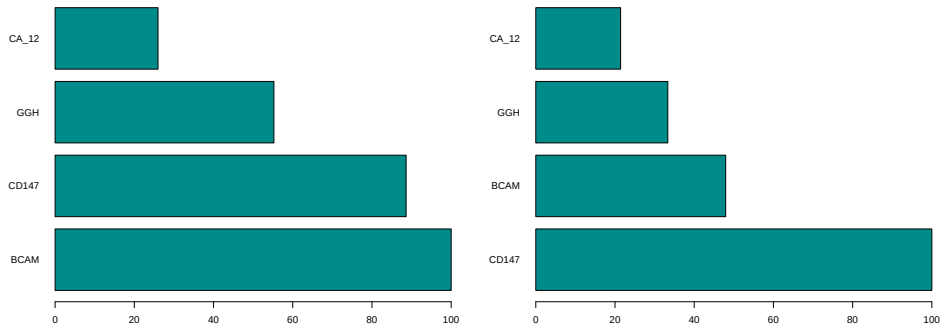


Figure 8: Left: benign or malignant, Right: basal or luminal

	Sensitivity	Specificity	Accuracy
benign vs. malignant	0.78	0.54	0.71
basal vs. luminal	0.73	1.00	0.90

Table 5: Results from random forest using leave-2-out-cross-validation.

4. Conclusion

LLR generally outperformed Random Forest for classification. LLR had strongest results with the basal vs. luminal classification, with the claudin-low classification dropping substantially in terms of sensitivity, and the benign vs. luminal classification more poorly in both sensitivity and specificity. We believe this is due to the benign vs. malignant and claudin-low data being very uneven, with only 5 positive observations in each classification.

Random Forest was outperformed by LLR in each terms of overall accuracy in each classification accept for benign vs. malignant. Additionally, it seemed that Random Forest heavily favored one class each time, with either one of sensitivity or specificity being very low. It would appear random forests was even more sensitive to the positive-negative spread of the classifications than LLR.

LLR and random forests both performed poorly at predicting benign or malignant classification on the second dataset; however, they performed exceptionally well at predicting basal or luminal classification. LLR was in fact able to perfectly classify basal or luminal in leave-2-out-cross-validation.

We obtained overlap between the important predictors selected through LLR and random forests. We are currently working to see how accurately these predictors classify the cell lines through additional methods such as Linear Discriminant Analysis and Naive Bayes.

While we have been able to attain high specificity and sensitivity, we are worried about over-fitting in our models. Cross-validation is prone to gathering noise and may not always pick reliable predictors when $p >> n$. We are currently looking into new methods such as estimation stability cross-validation which are meant to address this concern. [5]

5. Acknowledgements

- Initial statistical analysis was funded by NIH Grant R15 CA164929.
- Spencer Bowen was supported by the SFSU $(CM)^2$ NSF GK-12 program (Grant DGE-0841164) throughout the creation of this poster and report.
- The mass spectrometer used for measurement of glycoprotein data was purchased with funds from NSF-MRI Grant CHE-1228656.
- We would like to thank Dr. Bruce Macher and Dr. Robert Yen for collecting these datasets and providing them to us for analysis.

References

[1] Leo Breiman, *Random forests*, Machine Learning **45** (2001), no. 1, 5–32 (English).

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software **33** (2010), no. 1, 1–22.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[4] Andy Liaw and Matthew Wiener, *Classification and regression by randomforest*, R News **2** (2002), no. 3, 18–22.

[5] Chinghway Lim and Bin Yu, *Estimation stability with cross validation (escv)*, (2013).

[6] Ten-Yang Yen, Bruce A. Macher, Claudia A. McDonald, Chris Alleyne-Chin, and Leslie C. Timpe, *Glycoprotein profiles of human breast cells demonstrate a clear clustering of normal/benign versus malignant cell lines and basal versus luminal cell lines*, Journal of Proteome Research **11** (2012), no. 2, 656–667.