Simon Ng, Gregory Tom
Professor Meyer
BE 188
20 March, 2019

Modeling radiation exposure with blood gene expression from Paul et al.

Introduction/Motivation

In the modern age, the threat of nuclear attacks necessitates the ability to administer appropriate treatments for many thousands of individuals exposed to radiation. An important component of such a response is quantifying the level of radiation an individual has received. The current gold standard for radiation level prediction is the dicentric assay. The assay capitalizes on the fact that radiation causes chromosomes to fuse and measures the number of chromosomes with two centromeres. While this method is reliable and the results are constant over weeks, it has a 1-2 day turnaround and is difficult to automate [1].

In the event of mass triage following a nuclear attack, efficiency of treatment is of the utmost priority. Thus, researchers have been searching for a better method of accomplishing this task than the dicentric assay. We explore a similar approach to Paul et al. by using gene expression levels from peripheral blood cells in order to predict radiation exposure levels.

Problem Definition

Radiation exposure has been shown to affect gene expression levels in peripheral blood cells [2]. Peripheral blood cells are especially useful for this type of study because they can easily be biopsied and their gene expression remains altered for several days after exposure [3]. This presents an opportunity to use peripheral blood gene expression measurements for the purpose of predicting continuous radiation exposure levels. This method would present an advantage over the use of the dicentric assay because rapid high throughput screening of expression signatures is possible with biochips [4].

However, what complicates this process is that there are thousands of genes that can be measured and an algorithm needs to find which genes are most important for accurate prediction. Though radiation can affect each gene differently, a holistic analysis of all expression levels should be able to give an accurate prediction of exposure levels. Paul et al. addressed this problem by implementing a 3 nearest neighbors algorithm that classifies patients into one of five radiation exposure levels.

The goal of our project was to implement a machine learning algorithm from class to accomplish the same purpose of predicting. We also wanted to identify the biological function of the genes affected most by radiation in order see which specific biological functions radiation targets.

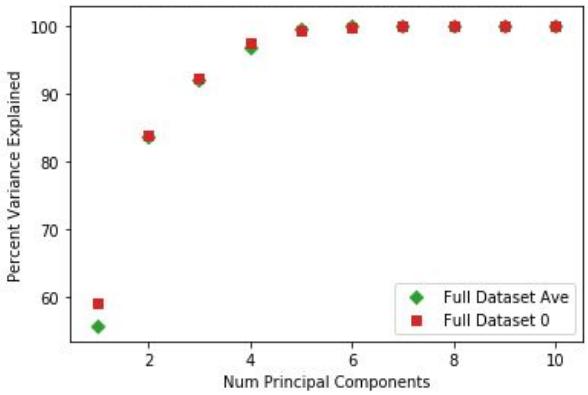

Fig 1. Comparison of percent variance explained using dataset with missing values filled using averaging or zeros. Filling with zeros yields a small advantage at fewer principle components.

Methods

Model Selection

We chose to use the Partial Least Squares Regression (PLSR) algorithm. This is because the data set we obtained from Paul et al. consisted of 25 patients and 41,000 gene expression measurements. With many more variables than observations, the dimensionality reduction capabilities of PLSR were deemed important for interpreting the data. We used sci-kit's *sklearn.cross_decomposition.PLSRegression* function for this purpose.

Manipulating Data for Use

Each patient blood sample was exposed to one of five radiation levels: 0, 0.5, 2, 5, or 8 Gray and expression measurements were taken 48 hours after initial radiation exposure. The data set had missing gene expression values missing completely at random, so two approaches to filling in missing data seemed reasonable. The first option was to fill all missing values with zero. The second option was to average all the available values for the gene and set all the missing values to that average. The difference of this can be seen in Figure 1. While there was not a substantial difference in the variance explained (R2Y) value, filling in with zeros did yield slightly better results for fewer principle components. We built models using both methods, and they performed comparably. The averaging approach was much more computationally demanding, so all methods presented hereafter used the zeros method.

Secondly, in order to load the data into the PLSR function, we normalized the data by z-scoring. In order to avoid indeterminate values after z-scoring, we set all columns with a standard deviation of zero to all zeros.

Building the Model

PLSR was performed on the full 25 x 41,000 dataset both without cross validation

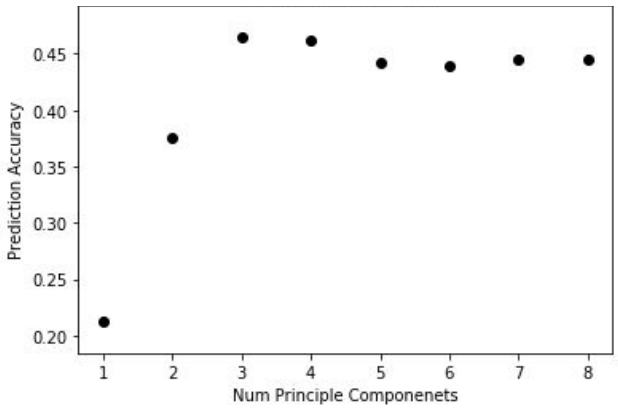

Fig 2. Prediction accuracy of full dataset

and with Leave One Out cross-validation. Building the model with Leave One Out yielded a maximum 47% Q2Y value as seen in Figure 2. While this was better than chance, we questioned whether a more effective model could be built. Based on the full dataset scores and loadings plots in Figure 3, we expected that the most important genes should be located either positively on PC 1 and negatively on PC 2, corresponding with high radiation exposure, or negatively on PC 1 and positively on PC 2, corresponding with low radiation exposure. Our next goal was to build an objective method to identify these important genes.

T-test Filter

In order to objectively select this subset of variables, we performed a series of t-test comparisons on the dataset. We compared each of the doses to each other dose to determine which genes exhibited significantly different expression across doses. A t-test was an ideal choice because our sample size was small, with only 5 samples per dosage. The 5 dosage levels yield 10 t-test comparisons for each gene. We chose to keep genes with 7 or more significant t-test comparisons. During cross-validation, the t-test filter was applied to the dataset after splitting. To increase computation speed, genes with fewer than 3 significant t-test comparisons were eliminated from the dataset prior to data-splitting. This was justified because splitting the training and test set removes one observation from one dosage level, affecting the 4 t-test comparisons involving that dosage level. The other 6 comparisons are between dosage levels completely unaffected by data-splitting. Thus, removing one observation from the dataset makes at most 4 additional significant comparisons. Genes with fewer than 3 significant comparisons prior to data-splitting cannot possibly become genes with 7 or more significant comparisons after data-splitting and therefore can be safely omitted. This method produced a 139 gene model.

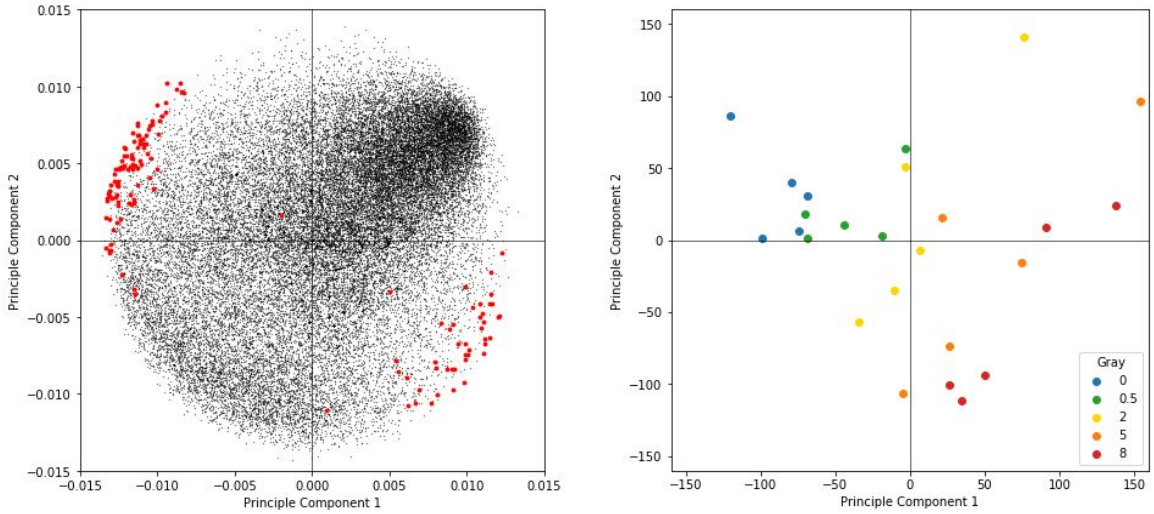

Fig 3. (left) Full dataset model loadings plot with significant genes colored in red as determined by t-test filter. (right) Full dataset model scores plot with labeled radiation dosage. Note that the observations roughly follow the line y = -x, crossing through the 2nd and 4th quadrants, similarly to the red genes in the loadings.

Biological Function Analysis

We used PANTHER [5] to identify the functions of genes identified as significant. We used Ensembl BioMart [6] to identify the Ensembl gene identifiers for Agilent probe IDs in the AGILENT WholeGenome probe microarray used by Paul et al. The Ensembl list was loaded into python using the pandas library and matched with Agilent probe IDs of the 139 significant genes, yielding a match for 127 genes. We input these Ensembl gene IDs into PANTHER to get a plot of biological function, the metric reported by Paul et al.

Results

The use of PLSR in order to model these predictions was an apt choice. According to Figure 1*,* the R2Y value on the full dataset was almost 100% at 6 principal components, which evidences a very low fitting error. In terms of the predictive capabilities, the final version of the model with cross-validation and the t-test filter performed very well. The model performed equally well from 6 to 8 principal components with a Q2Y value of 80% as seen in the left of Figure 4. The right of Figure 4 displays a graph of measured data vs. predicted data in order to show how well the model was able to predict. These results cannot be directly compared to the results of Paul et al. because their paper uses a classification technique; however, our model results seem to be at least comparable to, if not better than, their reported 72% classification accuracy. This is an intriguing result because it means that out of over 40,000 genes, only a very small subset is important for prediction purposes. It also suggests that radiation does not alter all genes in peripheral blood cells but only a select few.

In the reduced dataset scores and loadings plots in Figure 5, the observations are distributed along PC1 from low to high dosage going from left to right with little meaningful information encoded by PC2 as both 0 and 8 Gray are located positively on PC2. The genes in the loadings plot are divided in half along PC1. Genes negative on PC1 are likely downregulated by increased radiation exposure while genes positive on PC1 are likely upregulated. The t-test filter seems to select the line y = -x from the full dataset loadings in the left of Figure 3, along which the significant genes are located as PC1 for the reduced dataset.

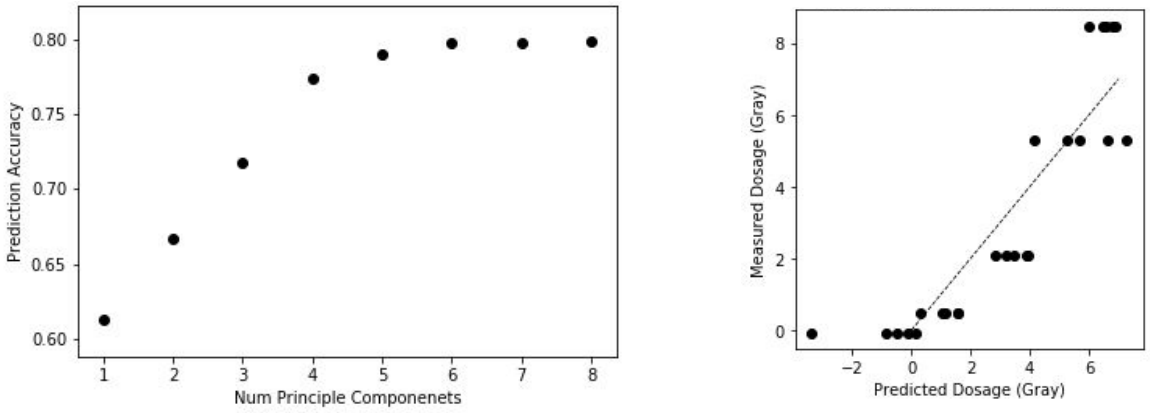

Fig 4. (left) Prediction accuracy of dataset limited to genes with 7 or more significant comparisons. (right) Measured versus predicted radiation doses for 8 principle components. Dashed line is ideal y = x line.

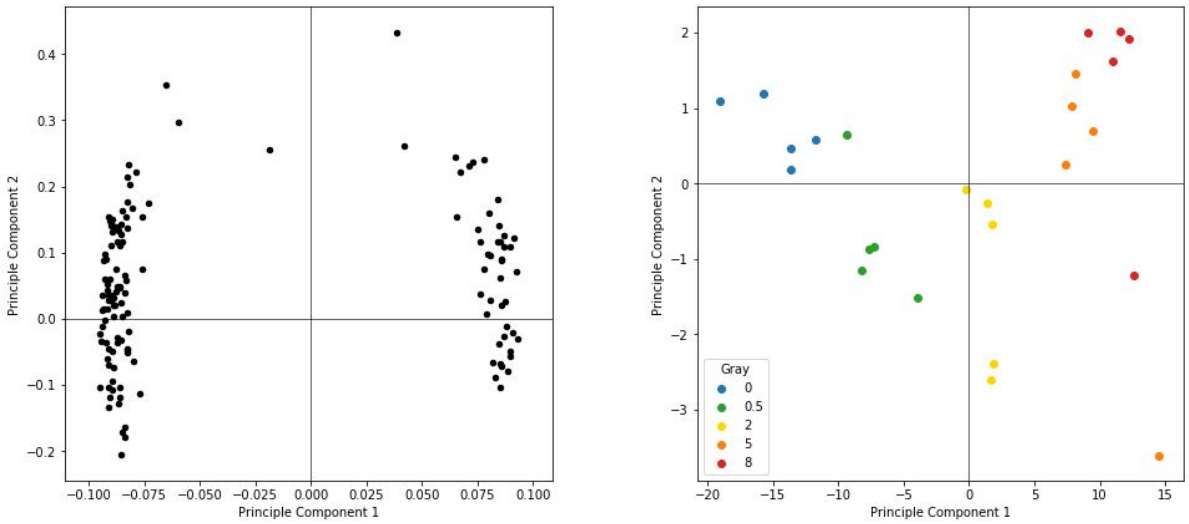

Fig 5. (left) Loadings plot for post-T-test filter model. (right) Scores plot for post-T-test filter model with labeled radiation dosage.

Furthermore, it is useful to know what processes the significant genes are associated with in order to better understand the physiological effects of radiation. The biological processes yielded by PANTHER do not match those reported by Paul et al. Their study identifies natural killer cell genes as a highly affected biological function, whereas our model varies greatly in types of gene function, as seen in Figure 6. Their identified genes were much more specific in function, whereas ours are more broad. This discrepancy could be due to the use of different modeling and data reduction techniques or more likely a misinterpretation of the methods used by Paul et al. to identify affected biological processes.

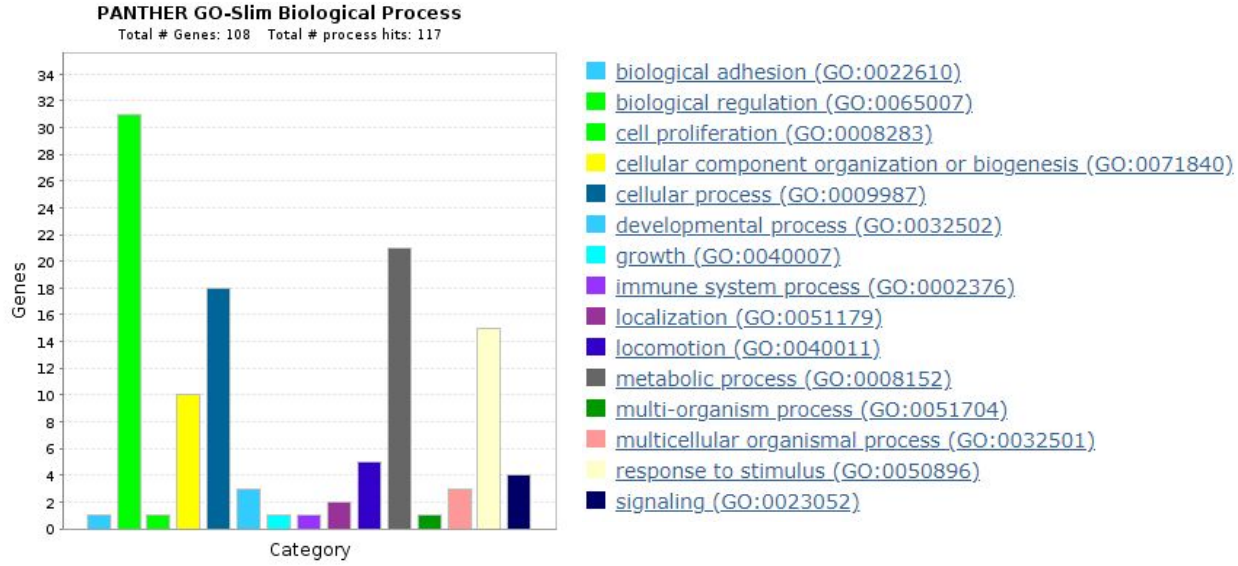

Fig 6. Biological processes of significant genes determined by PANTHER. Note the large variety of functions, which does not match the results from Paul et al.

Aside from this, we conclude that our model does explain variance well for the 48 hour time point after radiation exposure. It is uncertain whether the model will perform as well with data from other time points. Moreover, the data from the study was obtained from blood cells irradiated *ex vivo*. We do not expect *in vivo* gene expression to differ greatly from *ex vivo*, but it is a factor to consider. Therefore, further investigation is required to determine whether this model can be used in these other contexts. We are confident, however, that gene expression analysis can be a viable predictor of radiation exposure.

Division of Responsibility

Data entry was performed together, data formatting by Simon, percent variance, cross-validation, scores and loadings, and other initial code setup by Greg, t-test comparisons, PANTHER analysis, and code finalization by Simon, report outline and first draft by Greg, figure plots by Simon, final report editing together.

References

[1] Sproull, M. T., Camphausen, K. A. & Koblentz, G. D. Biodosimetry: A Future Tool for Medical Management of Radiological Emergencies. Heal. Secur. 15, 599–610 (2017).
[2] Amundson, S. A. et al. Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. Radiat. Res. 154, 342–6 (2000).
[3] Paul, S., Smilenov, L. B. & Amundson, S. A. Widespread decreased expression of immune function genes in human peripheral blood following radiation exposure. Radiat. Res. 180, 575–83 (2013).
[4] Brengues M, Paap B, Bittner M, Amundson S, Seligmann B, Korn R, et al. Biodosimetry on small blood volume using gene expression assay. Health Phys 2010; 98:179–85.
[5] Thomas P.D. et al. Ensembl 2018, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D754–D761