Rachit Mukkamala
20.260 Final Project
May 8th, 2022

Proteomic Pathways and Signatures associated with Squamous Cell Lung Cancer Pathology

Background:

For the final project, I chose to analyze data from the paper *"Proteogenomic Landscape of Squamous Cell Lung Cancer"* by Stewart et al. (2019)[1]. This study aimed to conduct an integrated analysis of genomic, transcriptomic, and proteomic data collected from resected squamous cell carcinoma tumor samples to uncover detailed molecular mechanisms underlying SCC tumorigenesis and pathology. The study also aimed to uncover associations between molecular changes in the genome/transcriptome and clinical outcomes in order to identify new biological targets for SCC treatment.

The authors approached this problem of SCC multi-omic profiling by collecting N=108 SCC snap-frozen tumor tissues which were surgically resected from patients with stage I-III SCC of the lung. The authors also compiled all available clinical data and physician notes for these 108 patients for future data analysis. The researchers then conducted targeted exome sequencing/copy number analysis, RNA sequencing, and LC-MS/MS proteomic mass spectroscopy to collect genomic, transcriptomic, and proteomic data for each sample, respectively.

The authors clustered the samples into three groupings based upon protein expression and pathway enrichment: "Inflamed" (enriched for immune biology pathways), "Redox" (enriched for reduction-oxidation metabolic pathways), and "Mixed" (enriched for Wnt/stromal signaling). Parallel transcriptomic analysis yielded clusters that only partially matched with the proteomic clusters, reflecting the reality that transcriptomic signals are poor predictors of proteomic signals.

The authors hypothesized that the "Inflamed" cluster's enrichment for immune pathways might be reflected in the tumor's histology, so the authors performed immunohistochemistry assays for CD8 (T-cell), CD33 (myeloid cells), and CD20 (B-cell) immune markers on the tumor samples. IHC analysis revealed significant infiltration of Inflamed subtype tumors with myeloid immune cells, suggesting that immunosuppressive immune cells are a key aspect of SCC pathology. Further histological analysis revealed lymphoid aggregates that resembled tertiary lymph nodes (TLNs) and stained strongly for CD20 (high B-cell presence). Survival curve analysis revealed that presence of these TLNs were associated with improved cancer survival and overall outcomes.

Lastly, the authors filtered a list of genes that had highly correlated copy number, RNA expression, and protein expression data. The authors combined this list with siRNA knockdown data to conclusively identify three key SCC targets: PSAT1, TP63, and TFRC.

Summary of my Novel Analysis Approach:

In this manuscript, the authors noted that they were unable to find any significant associations between the biological DNA copy number/mRNA abundance/protein abundance data and the collected clinical/pathological outcome data. The goal of my analysis is to use a novel supervised machine learning approach based upon partial least-squares regression and classification (PLSR/PLS-DA) to identify proteins that are significant drivers of SCC pathology. My hypothesis is that each individual protein has a small contribution towards tumor outcomes that does not exceed the cutoff for significance set by the authors, but by building aggregate models and calculating feature importance I might be able to tease out high-level pathways and general trends involved in each clinical outcome. [2]

---

[1] Stewart, P.A., Welsh, E.A., Slebos, R.J.C. *et al.* Proteogenomic landscape of squamous cell lung cancer. *Nat Commun* **10,** 3578 (2019). https://doi.org/10.1038/s41467-019-11452-x

[2] All code, plots, results, protein/pathway lists, scores, etc. can be found in the attached Jupyter Notebook.

Methods and Results:

I began my analysis by cleaning the proteomics data matrix and converting it into a CSV format. I then loaded in all the clinical data sources, selected out the eight features that I was most interested in: Death(y/n), Recurrence(y/n), Cancer Stage, Grade Differentiation, Tertiary Lymph Node Score(0-3), Percent Tumor Cellularity, Percentage of Lymph Nodes with Positive Metastasis, and Percentage of CD20 Positive (B) cells. I then reformatted the compiled clinical data, encoded all the categorical variables into numerical values (ex: 1 for dead, 0 for alive, 2.5 for stage IIB cancer, etc.) and converted this clinical outcome matrix into CSV format.

The clinical outcome matrix had a small number of missing values, so I used KNN Imputation with k=1 to fill in those missing entries. In the proteomics matrix, most of the genes that were quantified had fewer than five missing values our of 108 samples. I filtered out all genes that had >60% of their values filled in, leaving me with 6,045 unique genes in the proteomics matrix.

For my first pass through the data, I chose to conduct unsupervised clustering, differential expression, and PCA analysis on just the proteomics matrix to explore the data's structure and attempt to replicate the clusters obtained by the authors. Since the tumor samples had no inherent classification structure within them, I did not have a way to cross-validate the ideal hyperparameters for imputation, so I just used the default value of k=5 to impute the proteome matrix using KNN Imputation. I then constructed an elbow plot using K-Means clustering and identified a clear bend at K=2, and visualization of the two clusters using PCA showed a very clear separation of the two clusters along PC1.

I used the Mann-Whitney U Test to identify differentially expressed genes that uniquely identified each cluster, and then conducted pathway enrichment analysis on each gene listing. One of the clusters had increased expression of mRNA splicing/processing proteins and decreased expression of secretory and extracellular proteins, whereas the second cluster had the exact opposite enrichment profile. Pathway enrichment of the genes with the highest and lowest loadings along PC1 confirmed this trend: genes with positive loadings were enriched for mRNA processing pathways and genes with negative loadings were enriched for extracellular secretory pathways. Interestingly, this clustering pattern of secretion vs. mRNA processing did not match the clustering paradigm of inflammation, redox metabolism, and Wnt signaling that was observed by the authors.

Since the tumor clusters were well defined, I used them as a proxy classification variable to assess KNN Imputation performance. I used sklearn's GridSearchCV function to test multiple values of K and evaluated imputation quality by assessing the ability of a Random Forest Classifier fitted on the imputed data to predict tumor clusters. Using this approach, I determined that K=6 had the best performance and re-imputed the protein abundance matrix.

My first attempts at fitting a PLSR model on the entire data yielded poor results, with Q^2 values consistently in the negative range. I then attempted to only fit the model on one "Y" variable at a time, but this still led to sever overfitting. To combat overfitting, I incorporated Ridge, Lasso, and ElasticNet regularization (with cross-validation optimized alphas) into the fitting pipeline as an initial feature selection step, but none of these algorithms yielded meaningful improvements.

I then attempted to fit a variety of other supervised models, such as Random Forest Classifiers/Regressors, Logistic Regression (with PCA transformation), and Principal Components Regression. Unfortunately, all of these models also yielded poor accuracies with heavy overfitting. Given how the authors also were unable to meaningfully find associations between proteomic and clinical data, I was ready to move on from supervised modeling. However, as a last-ditch attempt, I tried to conduct feature selection by selecting all genes which had a magnitude of correlation with the outcome variable that was greater than or equal to 0.15, and this approach surprisingly yielded excellent results (see accuracies in next section)!

After settling on my correlation feature selection approach, I tried three models for classification of categorical data: PLSDA, Random Forests, and Logistic Regression. All three models performed equally well, so since I was already using PLSR for numerical data I decided to use PLSDA for classification to maintain consistency. I then proceeded to write abstract helper functions to reduce the amount of copied code necessary to fit and analyze eight separate PLS models. I also subclassed the sklearn base module and created a custom PLSClassifier() sklearn model so that I could directly pass PLSDA into sklearn pipelines. My modeling pipeline was as follows:

1. I conducted feature selection by filtering out all proteins with correlations to the outcome variable that are less than 0.15 (or 0.20 for a model that is not performing well). I then passed the feature-selected proteomics and clinical data to a custom model-fitting function.
2. In the model fitter, I used GridSearchCV to find the optimal number of components for the PLS model.
3. I created the model, split the predictor and outcome variables into train and test sets (training size = 80%), and Z-scaled the train and test predictor matrices.
4. I calculated the performance of the model on the training data (accuracy for classification, R^2 for regression).
5. I passed the X and Y matrices to a helper cv_score() function, which ran 5-fold cross validation on the data and computed the test performance metrics.
6. I returned the training model and used a custom VIP_Scores() function to compute the VIP scores for all proteins that were initially selected.
7. I split the proteins into groups based on whether their model coefficient was positive or negative. I selected the proteins with the top 40 VIP scores from each group and performed pathway enrichment on each of these final protein shortlists.

I repeated these seven steps on all eight clinical outcome variables, one by one. For certain clinical variables with multiple categories (such as cancer stage), I fit both a PLSR and a PLSDA model to the data, especially since the categories also implicitly encoded severity level (ex: a Stage IIIB tumor is more severe than a Stage IIA tumor), making regression just as appropriate as classification in these circumstances.

Detailed information regarding pathways and genes associated with each outcome variable is listed below. In general, when running the model-fitting pipeline multiple times, I noticed that the pathway enrichments experienced drastic changes between runs for certain variables (ex: Death, Tumor Cellularity) yet stayed highly consistent for others (ex: Recurrence, Percent CD20). Across all eight variables, I noticed that the VIP scores were very close together and nearly all VIP scores, even the largest ones, were less than or equal to one. This result confirms my initial hypothesis that changes in clinical and histological outcomes among SCC patients are determined by small, pathway-level changes in multiple proteins which cumulatively cause large changes in outcomes.

Summary of Biological Insights and Next Experimental Steps

My first novel biological insight was the identification of two new SCC tumor subtype clusters based on protein abundance: the mRNA processing cluster and the secretory cluster. While my clustering scheme did not match the clusters obtained by Stewart et al., I believe that my clusters are valid (and potentially more informative) as evidenced by their clear separation along PC1, the clear negative correlations between the clusters, and the ease with which a Random Forest classifier was able to predict each tumor's cluster assignment (see the section above regarding imputer optimization, test accuracy > 99%). One experimental next step that is needed is to reconcile my clustering approach with that of Stewart et al. and determine which clustering is more accurate. The tumor subtype clusters can also be experimentally validated through analysis of a new set of SCC samples to determine whether the same patterns emerge in the new samples.

My largest new biological insight is the identification of key pathways and proteins involved in eight important clinical features of SCC using supervised PLS models. These results are summarized below:

- **Death:** PLSDA was highly capable of predicting patient death, with a test data accuracy of 91.7%. The top five features associated with patient death were B4GALT4 , RPS6KA1, PAK4, SMAD5, and CAB39. Pathway enrichment for death-associated pathways was unreliable and changed significantly between model reruns. Thus, it is more informative to focus on the list of key death-associated genes (listed in the ipynb code).

- **Recurrence:** PLSDA was highly capable of predicting tumor recurrence, with a test data accuracy of 86.1%. The top five features associated with recurrence were LARP4, LY6D, GBA, COL6A1, and SERPINB7. Pathway enrichment showed that recurrence was consistently positively associated with DNA replication pathways negatively associated with lysosomal pathways.

- **Tertiary Lymph Nodes:** PLSDA performed better than PLSR at predicting a patient's TLN score, with a test accuracy of 59.2% and less overfitting. The top five features associated with higher TLN scores were CDK2, PELP1, MAP1A, PNPT1, and PFKL. I was unable to perform pathway enrichment since TLN Scores had four possible classes (0-3). The Stewart et al. paper showed that tertiary lymph nodes were positively associated with patient survival, meaning that these proteins could be associated with increased survival.

- **Differentiation:** PLSR performed better than PLSDA at predicting tumor differentiation, with a test Q^2 of 0.536. The top five features associated with tumor differentiation were ATRNL1, PKLR, TOP2B, SGPP1, and CGNL1, and all five of these genes had VIP scores > 1. Pathway enrichment was not reliable, but ER-related pathways were positively associated with differentiation, whereas blood clotting and apoptosis pathways were negatively associated with differentiation. Apoptosis has appeared as a positive and a negative associated pathway during multiple model fits.

- **Stage:** PLSR performed better than PLSDA at predicting tumor stage, with a test Q^2 of 0.516. The top five features associated with tumor stage were IFIH1, POLR3E, PADI3, MLST8, and VAV1. Positive pathway enrichment consistently showed that cancer stage was positively associated with mitochondrial/cell respiration pathways, but negative enrichment was not reliable.

- **Percentage of Lymph Nodes with Positive Metastasis:** Unfortunately, PLSR was unable to properly fit the data for this variable, as evidenced by a test Q^2 of 0.045. Thus, I decided to drop this variable from my analysis.

- **Percentage of CD20 Positive Cells:** PLSR modeled tumor CD20+ percentage moderately well, with a test data Q^2 of 0.411. The top five features associated with tumor B-cell abundance were CCDC88A, RAB9A, HSPA1L, FAM120C, AGA, and STRIP1. Positive pathway enrichment consistently showed that tumor CD20 percentage was positively associated with ribosomal pathways, but negative enrichment was less reliable as it showed negative associations with vesicle formation but occasionally ribosomal pathways as well.

- **Percent Tumor Cellularity:** PLSR modeled tumor cellularity reasonably well with a test data Q^2 of 0.513. The top five genes associated with tumor cellularity were TUBG1, RNF112, TBCB, FAM114A2, and KIAA0100, and all five of these proteins had VIP scores > 1. Pathway enrichment, however, was not reliable, as the only pathway that was enriched in both positive and negative features was the cytoplasm.

The next experimental steps for this data would be to validate these findings in vivo using mouse models. The procedure for these mice studies would be to pick one clinical variable, generate a CRISPR knockout of one of the top 10 genes associated with that clinical feature in SCC mice, and assess changes in that

clinical variable between the knockout group and SCC mice without any genes knocked out. It would likely be preferable to focus these experimental studies on clinical variables with either a high prediction accuracy (ex: Death, Recurrence) or a highly consistent pathway enrichment pattern (Stage, Recurrence, % CD20).

Additionally, given that B-cell infiltration is associated with changes in ribosomal pathways and cancer stage is associated with mitochondrial pathways, it would be very valuable to isolate tumors with high B-cell counts or a severe-stage tumor sample and analyze changes in mitochondria and ribosome morphology, respectively, using an electron microscope. Lastly, since we are now aware that SCC tumors involve a high degree of immune cell infiltration, it would be highly beneficial to conduct single-cell RNA sequencing analysis of SCC tumors to separate cells based on their cell type and analyze each individual cell type's unique transcriptomic signatures.

Conclusion

In this project, I report an analysis of SCC proteomics data using both unsupervised and supervised machine learning algorithms. My unsupervised analysis showed that SCC tumors can be separated into two distinct subtypes, one of which is characterized by high expression of mRNA processing proteins/low expression of secretory proteins, and the other which has the exact opposite signature.

My supervised analysis revealed that it is indeed possible to identify associations between tumor protein abundances and macroscopic clinical/histological outcomes. The key to my approach was that it used supervised PLS modeling to combine small changes in individual protein abundances into larger latent variables, which were then easily able to separate out samples by their clinical features. The most interesting findings from my supervised modeling were that: (1) B-cell infiltration is associated with increased ribosomal pathway expression, (2) higher tumor stage/severity is associated with increased mitochondrial/cell respiration protein expression, and (3) tumor recurrence is associated with increased abundance of DNA replication proteins and decreased abundance of lysosomal proteins. Apart from these large-scale trends, my analysis yielded rank-ordered lists of key proteins involved in seven key clinical variables which provides a very convenient starting point for researchers looking to validate new targets for SCC therapies.

One major lesson I learned from this project is the importance of conducting proper feature selection before fitting a machine learning model. Feature selection is a powerful way to eliminate noise that can ruin the predictive power of a model by encouraging overfitting. Before I conducted proper feature selection, my model's accuracy was so low that I would have been better off predicting the mean, yet once I ran correlation-based feature selection, I was able to boost my testing accuracies to >90% in some cases! Another important nuance to feature selection is that not all feature selection methods are equal, and the choice of feature selection approach is highly dependent on the dataset. For instance, although regularization is usually a solid choice for feature selection, it actually made the performance of my models worse in this particular dataset. Along the same limb, just because correlation-based selection worked so well this time doesn't mean that it will do the same always.

The second major lesson I learned is the power of well-trained supervised learning models to identify big-picture patterns and trends in the data by combining multiple predictor variables together. On their own, each individual protein in this dataset was an extremely weak predictor of clinical outcomes, since none of the proteins had a correlation magnitude higher than 0.3. This is the pitfall that Stewart and colleagues faced: since they looked for correlations between individual proteins/genes and clinical variables, they couldn't make any significant predictions. However, PLSR/PLSDA and ensemble-based Random Forest classifiers were able to combine the small effects of each variable together into a large-scale model that clearly separates apart tumors from different clinical classes. This is probably the largest and most important lesson that I took away from this project (and this course in general), and I look forward to continuing taking advantage of unbiased, algorithm-based feature importance calculation in my future big data/omics analysis projects.