Mutation Profile to Predict Tumor Stage in Lung Adenocarcinoma

Stanford | ENGINEERING
Mechanical Engineering

CS229a Introduction to Machine Learning Poster Session
Calvin Kuo: calvink@Stanford.edu

December 12, 2017
Project Category: Life Sciences

ABSTRACT

Lung adenocarcinoma is among the most common and deadliest cancers in the United States, accounting for an estimated 120,000 new cases every year, and responsible for nearly 10% of cancer deaths[1]. Because of its prevalence and high mortality rate, it is important to understand the histological progression for more targeted treatment and prognosis. Previous whole genome or whole exome sequencing of lung adenocarcinoma tumors have identified several genes that are significantly mutated in this cancer, and have provided insight into affected pathways and targets for treatments[2,3,4]. However, the type of treatment and prognosis depend on the severity of the cancer and its risk of spreading (metastasis). In this work, we found that we could differentiate early stage and late stage lung adenocarcinoma tumors with 60% accuracy using mutation effects in key genes. Feature reduction methods were key to reduce overfitting in training.

DATA PRE-PROCESSING

Data Source
- Data were obtained from a public repository (cbioportal.org)
- Two datasets containing 501[2] and 159[3] lung adenocarcinoma samples
- Tumor stage by physician, mutations by genome or exome sequencing
- 501 sample set used for training, 159 sample set used for testing.

Tumor Stage
- Size (T), Lymph node involvement (N), and metastasis (M)
- Stage 1 tumors are localized to lung, higher stages have invasive spread
- Labeled stage 1 tumors, and higher tumors for binary classification

Mutations
- Over 150,000 mutations occur throughout the genome in datasets
- Mutations in genes coding proteins are likely to have greatest effect
  - 36 genes previously identified as significantly mutated
  - Also included 32 genes with mutations in >25% of training samples
- Pruned to 11,943 mutations over 60 genes of interest

Feature Sets
- (1) Number of mutations total in 60 genes
- (2) Number of mutations in each gene
- (3) Number of mutation types in 60 genes (describes how mutation occurs)
- (4) Number of mutation effects in 60 genes (describes effect on amino acid coding)

Figure 1: Stages of Lung Adenocarcinoma

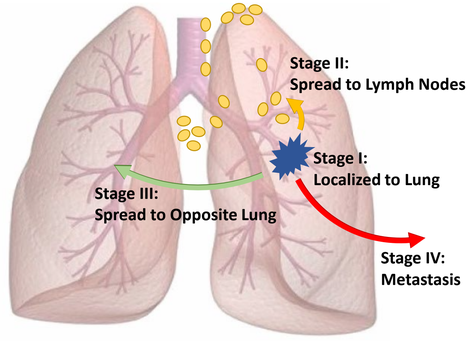
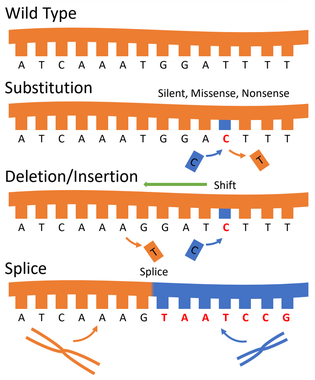


Figure 2: Mutation Types and Effects



METHODS

Initial Tests
- Used simple machine learning techniques on individual feature sets for binary classification
  - Logistic Regression
  - Naïve-Bayes (multinomial event model)
  - Support Vector Machine (linear, radial basis, sigmoid, and polynomial kernel)
- Method accuracy determined by % correctly labeled

Feature Space Reduction
- Feature sets (2) and (4) were most promising, with greatest training set accuracy (66%, 56%)
- Created new feature set representing number of mutation effects in each gene
- Large feature set required pruning to prevent over-fitting
  - Naïve-Bayes (NB) to determine most predictive features (top 25)
  - Principal Component Analysis (PCA) on training set features (>90% variance)

Machine Learning Techniques
- Support Vector Machine (SVM) with radial basis kernel
  - Forward Feature Selection (FFS) with 10-fold cross validation
- 3-layer Neural Network (NN)
  - 15 or 50-node hidden layer with sigmoid activations and log-loss cost

Figure 3: Processing Pipeline

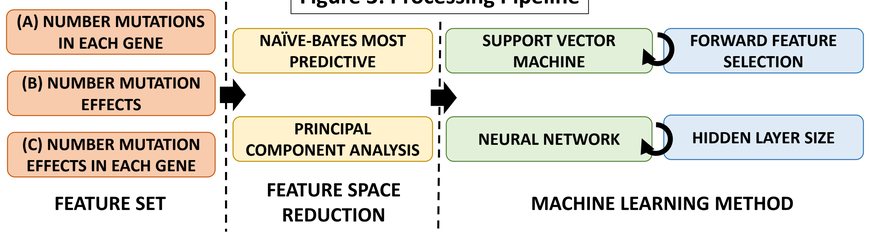
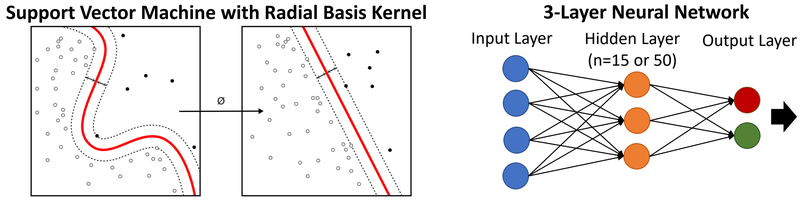
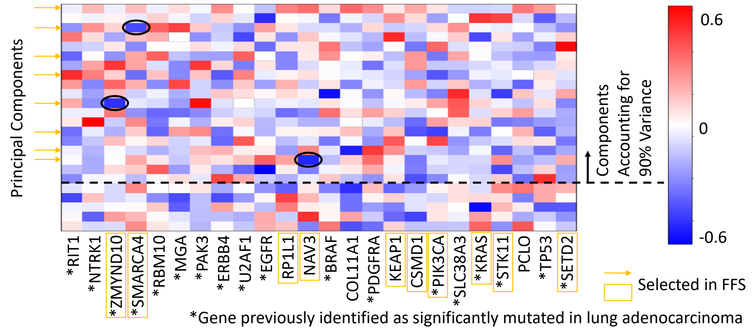


Figure 4: Machine Learning Methods

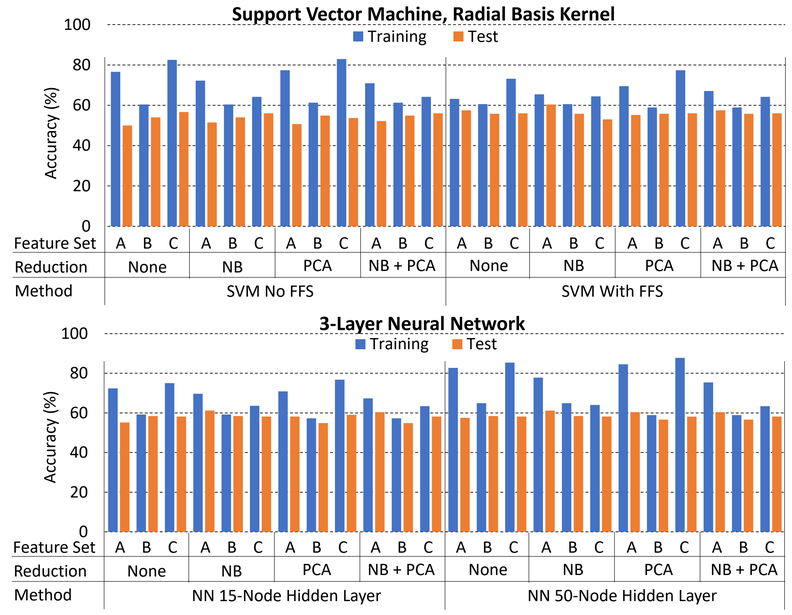


RESULTS

Figure 5: Principal Components for Number of Mutations in Each Gene Top 25 Most Predictive Genes from Naïve-Bayes



RESULTS



Conclusions

Conclusions
- Feature reduction techniques are necessary to prevent overfitting
- Selected features are severe mutations (splices, shifts) in genes associated with cell proliferation (ZMYND10, SMARCA4)
- Test set accuracy was generally <60%
- Tried redistributing training and test set data to homogenize two sets
  - Did not improve test set accuracy performance
  - Did give better agreement between training set accuracy and test set accuracy
- Tried SVM forward feature selection using test set accuracy as cost metric
  - Was able to obtain 81% training set accuracy and 75% test set accuracy
  - Can be considered upper bound on expected performance
  - But defeats the purpose of test set for unbiased validation of classifier

Future Work
- Mutation location in protein sequence dictates its effect on protein function
- Also, specific amino acid changes can result in greater functional disturbances
- Features that include these detailed information may be more predictive
- However, feature space would be massive (20 amino acids, proteins ~300-1000 codons long) and also require more advanced pruning techniques

Refrences

[1] National Cancer Institute, "Cancer Stat Facts: Lung and Bronchus Cancer" National Institutes of Health, https://seer.cancer.gov/statfacts/html/lungb.html.
[3] Ding, Li, et al. "Somatic mutations affect key pathways in lung adenocarcinoma." *Nature* 455.7216 (2008): 1069.
[3] Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." *Nature* 511.7511 (2014): 543.
[4] Bhattacharjee, Arindam, et al. "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." *Proceedings of the National Academy of Sciences* 98.24 (2001): 13790-13795.
[5] Campbell, Joshua D., et al. "Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas." *Nature genetics* 48.6 (2016): 607.