Machine Learning Classifier to Predict Liver and Pancreatic Carcinoma

Aparajita Choudhury (ac5901)
Anirudhan J Rajagopalan (ajr619)

Learning Problem and Motivation

- Predict Hepatocellular Carcinoma (HCC) and Pancreatic Adenocarcinoma (PAAD) using DNA methylation and protein expression.

- DNA methylation profile and protein expression are known to be different between cancer and normal cells.

- Potential to be used as effective biomarker indicators for gene targeted treatment.

Existing Research

- Existing work:
  - DNA methylation profile individually
  - Gene expression individually
  - Comparison between DNA methylation and gene expression

- Not much has been done using protein expression in comparison to other biomarkers as an indicator of cancer.

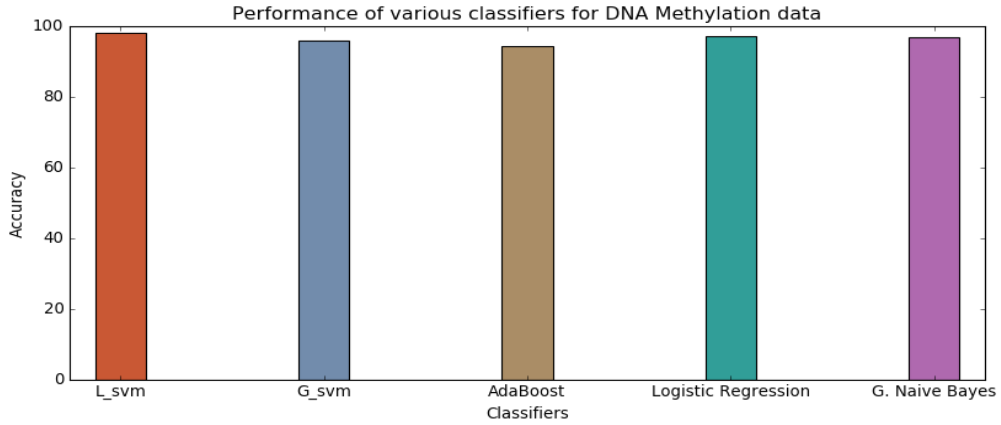
Dataset

- Data from TCGA for HCC and PAAD:
  - DNA methylation (Processed data: ~4GB)
    - 485577 features for both HCC and PAAD
    - 430 samples for HCC
    - 195 samples for PAAD
  - Protein expression (Processed data: ~1MB)
    - 205 overall features for HCC and PAAD
    - 184 samples for HCC
    - 123 samples for PAAD

Pipeline

- Imputation and normalization of data
- K-Means - with 2 and 4 centroid
- Models
  - AdaBoost - on reduced dimension
  - Linear and Gaussian SVM – on reduced dimension
  - Gaussian Naïve Bayes – on full feature set and reduced dimension
  - Logistic Regression – on full feature set and reduced dimension

Comparison of models



Conclusion

- DNA methylation – best model: Linear SVM with 98.09% accuracy
- Protein expression – best model: Gaussian SVM with 93.51% accuracy
- Tradeoff: Accuracy vs Resource Utilization
  - DNA methylation: Higher accuracy – Running Time of ~4 hours – Requires ~4GB data
  - Protein expression: Relatively lower accuracy – Running Time of ~15 minutes – Requires ~1MB data

References

- John Hayward, Sergio A. Alvarez, Carolina Ruiz, Mary Sullivan, Jennifer Tseng, Giles Whalen – "Machine learning of clinical performance in a pancreatic cancer database", Artificial Intelligence in Medicine 49 (2010) 187–195
- Krzysztof Pawlowski, Frederic Pio, Zhi-Liang Chu, John C. Reed and Adam Godzik – "PAAD – a new protein domain associated with apoptosis, cancer and autoimmune diseases", TRENDS in Biochemical Sciences Vol.26 No.2 February 2001
- PeiWei Zhang, Lei Chen, Tao Huang , Ning Zhang , XiangYin Kong , YuDong Cai – "Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles", Published: March 30, 2015 DOI: 10.1371/journal.pone.0123147
- Vivek Jain, Weizhuang Zhou, Yifei Men – "Machine Learning Classification of Kidney and Lung Cancer Types", Stanford University
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
- Yuan Yuan et al. – "Assessing the clinical utility of cancer genomic and proteomic data across tumor types", Nature Biotechnology 32, 644–652 (2014) doi:10.1038/nbt.2940
- Data: TCGA: https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm