Identifying Gene Expression Patterns Associated with Increased Survival Time Across Cancers

Introduction

Aside from predicting drug efficacy, genetic biomarkers can also function as important tools for cancer treatment. For example, by examining gene expression differences in tumors from different breast cancer patients and comparing their survival, it may be possible to learn what genetic differences lead to better breast cancer prognosis. However, performing this analysis *across* different types of cancers will hopefully identify genetic differences that contribute most to cancer patient survivorship in general. Thus, if multiple cancers can be shown to function similarly at a genetic level, then it may be possible to repurpose a drug therapy developed for one cancer to another.

Unfortunately, this is a complex task. Humans have over 20,000 genes. Simultaneously examining gene expression in this many genes for thousands of patients proves to be challenging. Therefore, we propose a machine learning approach to handle this complexity. In this project, we will apply supervised machine learning methods to publicly available cancer gene expression data to learn genes expression patterns that are associated with increased survival across multiple cancers (if any such patterns exist). Our goal is to identify genes that lead to increased survival times across multiple cancers.

Data

The Cancer Genome Atlas (TCGA) is a large project aimed to catalog genetic mutations that are responsible for a variety of cancers by sequencing tumors and measuring gene expression from patients in both tumor and normal tissue [1]. It involves multiple research institutions and is managed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) which are part of the National Institutes of Health (NIH).

So far, TCGA has collected data from thousands of cancer patients spanning thirty-four cancer types. We will analyze gene expression and clinical data for four cancer types: breast cancer, ovarian cancer, uterine cancer and cervical cancer. While pan-cancer research has gained momentum [2], researchers have not yet analyzed genetic commonalities across this combination of tumor types in relation to patient survival. Our total data will comprise 2,041 tumor samples and expression data from 20,530 genes [Figure 1]. Data values are real numbers representing the activity levels of genes in the samples and survival times for patients.

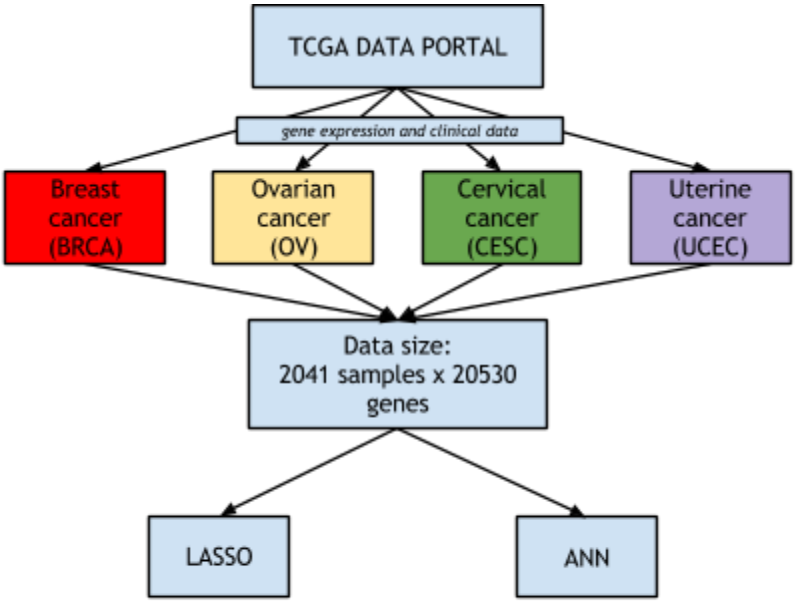

Figure 1 - Schematic design of our study data. Gene expression and clinical data will be downloaded from TCGA and integrated for implementation of our machine learning algorithms.

METHODOLOGY

We will use two primary algorithms, LASSO (Least Absolute Shrinkage and Selection Operator) and ANN (Artificial Neural Network) to complete our project objective. The first stage of our project will include model building with each algorithm. Next, we will select a subgroup of genes that best serve as predictors of survival across all tumor types. Finally, we will evaluate the performance of our machine learning methods by completing a literature search of the top ranked genes' annotations to confirm biological significance [3].

LASSO logistic regression builds a particularly efficient model of features, using only the variables that are the most informative [4,5]. It is a popular technique for selecting a sparse set of predictors in biological datasets (e.g. identifying the smallest set of genes that reliably predict if an individual would benefit from a particular therapy). LASSO optimization is similar to normal logistic regression, but it tends to reduce many of the feature coefficients to 0, leaving a relatively small set of features that are best able to predict an example's class. This ultimately removes redundant features and leads to a model that is easier to interpret than some other approaches. This model could be extremely useful for problems such as ours in identifying a smaller subset of genes that best predict survival.

As the name suggests, ANNs are made up of a network of interconnected nodes that were first adopted for mathematical modelling in 1943 [6]. The nodes are linked to each other by weighted connections, and we will use back propagation (BP) learning to construct our network [7]. BP is implemented in two cycles; a forward generation of output followed by backwards error propagation. The weighted connections are updated during the second cycle and this is carried out until the error in

the network is minimized. Lastly, we will compare the results of our algorithms with baseline methods to assess performance of our trained models.

MILESTONE

By February 17, we will have completed our code for LASSO and ANN and performed initial runs on a sample of our training data. The three major steps after this will be implementing our code on the entire data, comparing the results generated by our two models with other baseline methods, and functionally analyzing our genes of interest.

REFERENCES

1. Cancer Genome Atlas Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216), 1061-1068.
2. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J.M. (2013), The Cancer Genome Atlas Pan-Cancer analysis project, *Nature genetics,* **45**(10), 1113-1120.
3. The Gene Ontology Consortium. (2005). Gene ontology: tool for the unification of biology. *Nature Genetics*,**25**(1):25-9.
4. Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., and Zhang, H. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC bioinformatics*, **14**(1), 198.
5. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58**(1), 267-288.
6. Mcculloch WS, Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics,* **5***, 115–133*.
7. Abraham, Ajith. (2005), Artificial Neural Networks, *The Handbook of Measuring System Design,* John Wiley and Sons, New York.