Predicting the Clinical Outcome of Papillary Thyroid Carcinoma Patients through Publicly Available Omics Data

Kun-Hsing Yu, Luke Pappas, Michael Fitzpatrick, and Jessica Kung

Progress to Date:

1. Literature Review

We reviewed relevant literature on thyroid cancer and two papers on network-based classification approaches the TAs pointed out to us. We will integrate the network-based methods and compare our findings with the literature.

2. Identified Data Sources

We explored available data sources, including The Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium (ICGC). Although ICGC has data from additional 15 patients from Saudi Arabia, we only have information on their somatic mutations and copy number variations. With an aim to reduce batch effects and variations due to different populations, we decided to focus on the TCGA data set.

3. Preprocessed Omics and Clinical Data

We obtained clinical, DNA methylation, RNA-sequencing (RNA-seq), and proteomics data from TCGA (n=494). We identified demographic information (age, gender, and ethnicity), tumor stage, survival status, and survival time from the clinical files. We also extracted the beta values of DNA methylation, normalized gene expression and protein expression levels from the database.

We randomly selected 70% of our patients as our training set, which we utilized to select the top features and to train the classifiers. For instance, we applied least absolute shrinkage and selection operator (LASSO) to select and train our survival classifiers. We evaluated the performance of the resulting classifiers using the held-out 30%.

Preliminary results:

Survival prediction:

We used a LASSO Cox survival model to fit the survival outcomes of patients in our training set. LASSO Cox model suits our purposes because it accounts for censored data and selects the top features from our huge feature space efficiently. We first tried to predict survival outcome using gene expression data (RNA-seq). On evaluation, we showed that our model divided the patients in the test set into two prognostic groups, with significantly different survival outcome (p value = 3.98e-07).

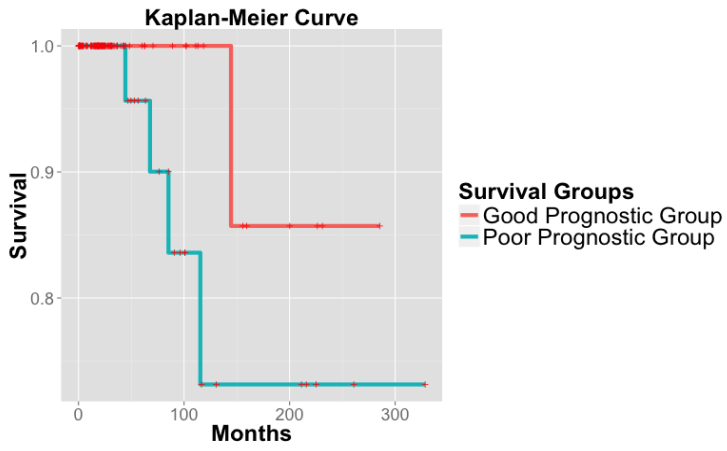

Figure 1. Survival outcome prediction using RNA-seq data. Our LASSO Cox model divided the patients in the test set into two prognostic groups, with significantly different survival outcome (p value = 3.98e-07).

Roadblocks Encountered:

TCGA consortium investigated the methylation status of >485,000 cytosine sites per patient, thus the resulting feature space and files are huge (12GB for all patients). We will try dimension reduction and feature selection methods to filter out uninformative sites and to reduce the computational cost.

Plans for Completing the Project:

- We will enhance our feature selection method using network-based approaches. We will also examine the biological functions of the top features through Gene Ontology (GO) analysis and Genomic Regions Enrichment of Annotations Tool (GREAT; for DNA methylation sites in non-coding genomic regions).
- We will build predictive models for tumor stage using the DNA methylation, gene expression, and protein expression data. This would illuminate the biological mechanisms involved in stage progression.
- We will investigate whether DNA methylation or proteomics information could also inform survival outcome of our patients.
- If there is time, we also hope to build prediction models for the three components of tumor stage (T: tumor, N: lymph node invasion, and M: metastasis) individually.
- In addition to presenting our analysis of the clinical outcomes, we also plan to build some web interfaces (charts, graphs, etc.) and tools to cleanly display our results. The extent to complexity of these visualizations and tools will largely depend on how much time we have after obtaining our results.