Medical Topic Modeling using dLDA on mRNA-Seq Multi-Cancer Datasets

Sarah Davis, Amira Aissiou, Luke Kumar, Laura Lee, Russ Greiner

Abstract

Cancer is an acquired disease involving genetic mutations and altered gene expression. Based on the idea that the altered expression of a single gene influences the gene expression of several other genes, we hypothesize that cancers can be characterized as a collection of patterns involving altered genes' expression. This project uses Latent Dirichlet Allocation (dLDA) borrowed from natural language processing to extract these patterns of altered gene expression, called "topics", from a large data set of solid tumor biopsies described by RNA-seq values. We use various machine learning model to make predictions on the clinical variables given the topic profile of a given tumor.

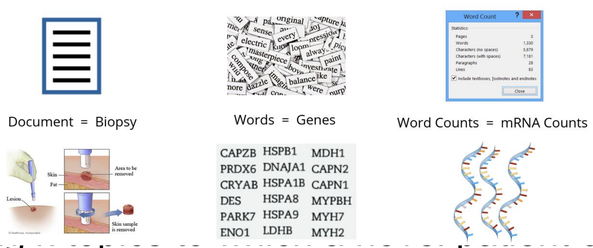
Methods

Data:

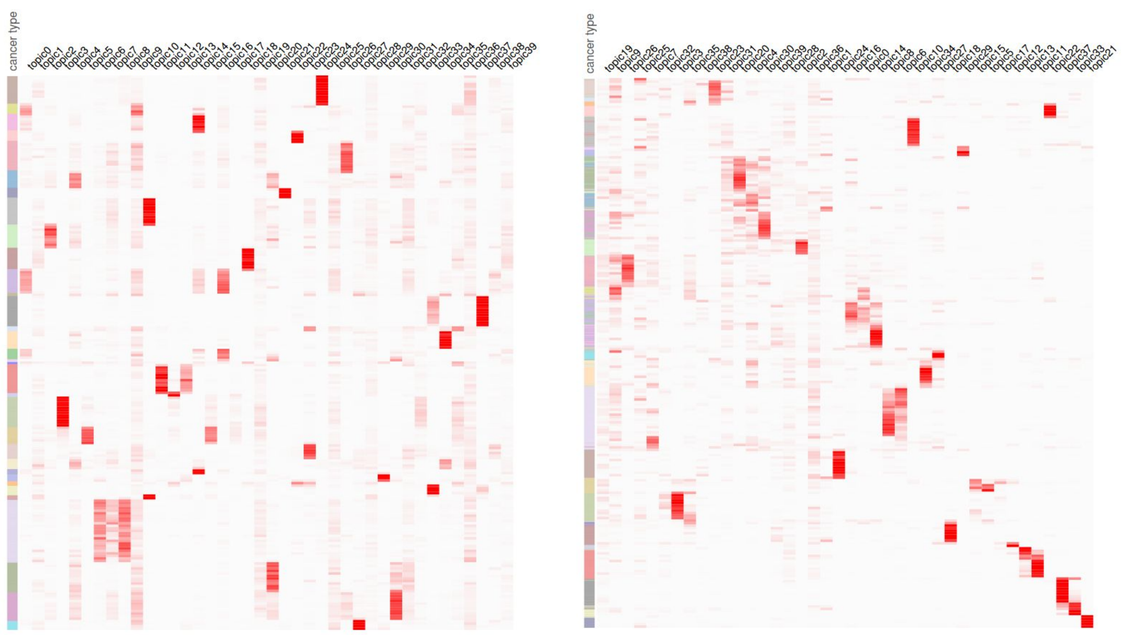
Our project uses 10,446 instances of The Cancer Genome Atlas (TCGA) RSEM-normalized transcriptomic data from 33 different cancer types and corresponding "pseudonormal" samples. These data include, for every tissue biopsy, the mRNA counts for each of 20531 genes.

Topic Modeling:

Latent Dirichlet Allocation (LDA) is a type of statistical modeling that uses the co-occurence of words to extract "topics" from a collection of documents. Luke Kumar, as part of his M.Sc. thesis, applied this method to biology using the following analogy:



Thereby generati ould have their membership calculated. Each patient was then represented by a *k*-tuple, that is their association with each of *k* topics.



Topic per Biopsy Heatmaps, where the rows are biopsies and the columns are topics. Left, the rows are sorted by cancer type. Right, rows and topics are clustered based on similarity.

We extend this model by applying it to multiple cancer types and using the *k*-tuples to predict patient clinical features such as cancerous biopsy vs. non, the type of cancer an individual has, and more.

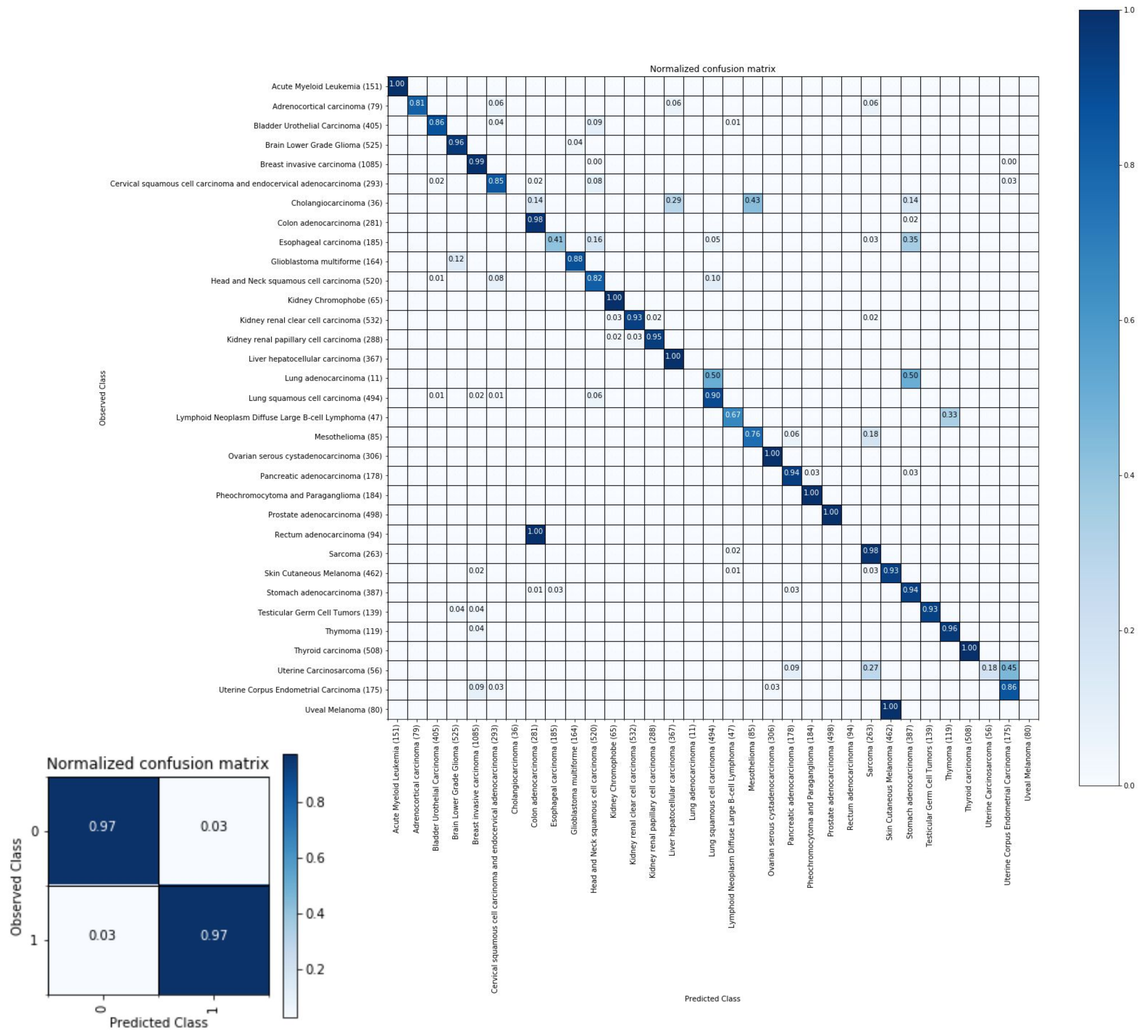
Predictions:

To test the validity of this dimensionality reduction, we trained Random Forest Regression (RFR), Logistic regression (LR), and Patient Specific Survival Prediction (PSSP) models to evaluate information retention.

Results

As noted beforehand, LDA can be used as a form of dimensionality reduction. We were able to represent any given patient as a 40-tuple rather than a 20531-tuple with present (but low) data loss. Because of this reduction in size, certain classes of models become available to use with these data, such as Cox-PH and PSSP.

We used these 40-element patient representations as input vectors for various statistical machine learning models. The following confusion matrices represent the accuracy with which logistic regression was able to discern two different clinical attributes.



Bottom left, accuracy of predicting whether a biopsy was cancerous or not. Top right, accuracy of multinomial logistic regression for predicting cancer type.

We were also able to predict survival time with a C-statistic of 0.75 based on this data, which is notably higher than what is currently found in the literature.

Discussion

Using dLDA for pattern extraction and dimensionality reduction over mRNA data from multiple different cancerous origins has proven successful, yielding promising results. Certain models (such as PCA) do similar dimensionality reduction, but lack any sort of explanation of the meaning of the topics. We believe that because LDA can offer that functionality, this tool can eventually be made clinically relevant.

Future Work

We will incorporate healthy tissue data into our current dataset from the Genome Tissue Expression (GTEx) database and generate a new set of topics. Adding healthy data will allow us to view the unifying features of all cancers contrasted by the non-diseased tissue expression, which may provide insight into the pathogenesis of the disease.

dLDA generates an ordered list of genes that are important to each topic, so the next step is to associate each topic with a phenotype using semantically meaningful gene sets.