Identification of DNA Methylation Markers using Feature Selection and Deep Learning

Nijhum Paul and Dr. Rick Jansen  |  Dept. of Public Health  
Aaron Huber and Dr. Rahul Gomes  |  Dept. of Computer Science

BACKGROUND

DNA METHYLATION AND CANCER
• DNA methylation is a process that can affect gene accessibility and therefore gene expression. There has been a focus in research on investigating methylation information to predict the relationship between specific gene methylation or expression and cancer. Methylation can affect genes that are associated with suppressing or contributing to tumor growth and progression, and as such, patterns in methylation can be used to improve cancer prediction accuracy.

DEEP LEARNING FOR BIOINFORMATICS
• Current tools used to gather DNA methylation data can now retrieve well over 850,000 unique methylation values per sample. As these datasets tend to be too large for a human to reasonably parse through by themselves, computational methods need to be used.
• Deep learning, a form of machine learning, is often used to make predictions and classifications based on data (e.g. allowing a computer to generally tell if a picture contains a cat or a dog). Previous DNA methylation research for cancer prediction has shown to be effective.

METHOD

DATASET
• The current study utilized breast cancer methylation-array data from The Cancer Genome Atlas (TCGA), specifically the TCGA-BRCA dataset. This dataset contains over 800 unique files containing methylation data, with each file representing a single sample.
• Each methylation-array file can contain around 450k different methylation markers, each of which may have a unique methylation beta value (ranging from 0 to 1) for a specific sample.
• For testing our process, we have selected a subset of this data with the following characteristics:
  20 total samples:
   10 primary solid tumor tissue samples
   10 solid tissue normal samples
  487177 unique markers (features) across all samples

DEEP LEARNING
• Our overall data processing procedure involved several stages (as shown in Figure 2). Data from the subset we selected first went through a stage of feature selection (via feature engineering processes), where data was filtered down to a more manageable size.
• Next, the data was split into three parts, based on samples: 12 samples went to training data, 4 to validation data, and 4 to testing data.
• Finally, we trained and tested a neural network in the deep learning stage of this process. The training and validation data was used to train the model, and the testing data was used to check the performance of the model as a result of training.

FEATURE ENGINEERING
• Each methylation marker is a different feature that must be considered in the deep learning process. Due to the large number of markers in this dataset, it is not feasible to run the entire dataset through the deep learning process in a reasonable amount of time, even with available computing resources.
• We are utilizing feature engineering to eliminate redundant or useless information from this dataset. Feature engineering involves the usage of different kinds of algorithms and statistical methods to identify pieces of data to remove. Across different contexts and applications, different combinations of these methods may be more effective under specific conditions.

• Utilizing our selected dataset, we have tested two feature engineering methods: the ANOVA F-Test and Random Forest. Utilized together, these reduced the total number of features from 487177 to 140645.

• Table 1 shows the names of top 20 features selected by ANOVA F-test. Feature Selection seems to have also improved the accuracy and sensitivity of the deep learning model, as shown in Figure 3.

DISCUSSION
• The feature selection techniques and current procedure utilized seem to be effective overall based on the subset of data used.
• The results seem to support the notion that deep learning methodologies for cancer prediction can be extended for uses in prediction of different types of cancer.
• A limitation of the current study was that the subset of data used was a relatively small proportion of the overall TCGA-BRCA dataset. This likely has limited the insights that can be gained from current evaluation metrics of our deep learning model.
• Moving forward, this model can be improved to make use of the full TCGA-BRCA dataset. As this will likely require significantly more computational power, plans are being made to utilize supercomputing resources at the University of Wisconsin - Eau Claire.

ACKNOWLEDGEMENTS
• This research was funded by NIH NDSU COBRE Center for Diagnostic and Therapeutic Strategies in Pancreatic Cancer (P20GM109024)

Table 1. Top 20 features based on ANOVA F-Test scores. Each feature refers to a genomic methylation site.