Identifying Acute Myeloid Leukemia using Machine Learning Methods on Selected DNA Methylation Data

Claire Hsu, Karunya Sethuraman

Abstract

Acute myeloid leukemia (AML) is a hematopoietic lineage cancer that presents as abnormal accelerated white blood cell growth in the bone marrow, producing large numbers of myeloid cells that do not function normally. Past experimental studies have shown that epigenetics affect cancer risk and proliferation via histone modifications, enzymatic modifications, DNA methylation, and other alterations. A model using novel combinations of classifier and feature selection methods can classify AML cases and identify the risk of malignant transformation of normal tissues. The information that this model uses to classify input DNA draws novel links between previously unstudied genes and AML.

Data was sourced from the Cancer Genome Atlas (TCGA-LAML) and non-malignant samples from five experiments found in the NCBI GEO database. The algorithm developed in this paper selected for significant features including CpG islands using multiple feature selection methods. Statistical analysis was done with a projection-based method in PCA, a supervised method in ridge regression, and a score rank-based method in T-test feature selection. Each selection algorithm was used to train a neural network and a Support Vector Machine (SVM). The accuracy of the neural net was found to be more dependent on the feature selection method used than the support vector machine was.

Different feature selection methods can highlight different patterns within methylation data, which differentiate between AML and non-AML cases. Our model utilizes CpG islands associated with genes which have been linked to AML and other cancers, as well as genes not previously linked, which could have implications on downstream factors increasing oncogenic potential. The application of these methods and models can lead to more selective identification of possible epigenetic drivers for cancer.