Name(s)

Karen Chung

Project Number

S0810

Project Title

Integrating Mathematical Modeling with Machine Learning for Cancer Driver Gene Identification

Abstract

Objectives

The identification of cancer driver genes is a critical component of precision oncology. Given the large feature space of The Cancer Genome Atlas (TCGA), which catalogs millions of somatic mutations observed in human tumors, machine learning techniques are ideally suited to driver gene identification. In existing models, however, the objective assessment of such machine learners is complicated by unexplained errors in the mutational data used to train the algorithms and by the absence of a perfect drivers list. This study employs mathematical modeling in tandem with machine learning processes to construct an objective and accurate classifier that identifies cancer driver genes.

Methods

In-silico dataset generation and machine learning training use the programming languages Go and Matlab, respectively; external datasets (TCGA, Cancer Genome Consensus, KEGG Database) are used later for classifier evaluation. A set of in-silico mutational data is generated by the stochastic simulation of a differential equations model of feedback-controlled cancer population dynamics. The synthetic dataset, validated through the assessment of mutational pattern distributions, trains a selected machine learning algorithm, producing a driver gene classifier.

Results

The gene classifier is shown to prioritize high-impact driver genes in four cancer types with >85% accuracy. Additionally, the quality of the ranked list of putative driver genes is validated through enrichment analysis on a list of generally accepted driver genes and biological pathway analysis. Top colorectal cancer driver genes from the classifier hold key roles in the PI3K-AKT and Wnt pathways, which have well-documented implications in carcinogenesis.

Conclusions

The interdisciplinary methodology developed here produces a more efficient and unbiased cancer driver gene classifier that can be utilized to identify henceforth unknown driver genes, providing insight essential for targeted cancer screening and treatment.

Summary Statement

I used mathematical modeling and simulation to generate synthetic data, which trained a machine learning classifier that accurately classifies and prioritizes cancer driver genes.

Help Received

I worked in the lab of Dr. Qing Nie at UC Irvine. My primary mentor in the lab was Dr. Seth Figueroa, who advised me on mathematical modeling and data simulation, and since Oct. 2018, I have been mentored by Matt Karikomi. Synthetic data generation was performed on the UCI cluster.