A Machine Learning Approach to Identifying Cancer Driver Genes

Karen Chung[1], *Seth Figueroa*[2], *Qing Nie*[3]

Northwood High School, Irvine, California[1], *Center for Mathematical and Computational Biology, University of California, Irvine*[2,3], *The NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine*[3]

Abstract

Recent advances in next generation sequencing has helped identify millions of somatic mutations in tumors from thousands of cancer patients. The vast collection and wide availability of DNA mutation data have encouraged the use of computational and machine learning approaches in identifying cancer driver gene candidates for use in precision oncology. However, the accurate identification of cancer driver genes versus the more frequent but incidental passenger genes remains a difficult task. The mutation data used to train the machine learning algorithm is large and complex; it is prone to errors and imbalanced in classes. These caveats and the lack of a gold standard driver gene list make it difficult to objectively assess and validate the accuracy of various computational methods. This study aims to compare choices of feature selection, data preparation, and machine learning methods in order to gain better insight on classification of driver and passenger genes. To address the previously stated problems with standard datasets of human mutational profiles, we develop an in-sillico training set of somatic mutation data. We simultate the accumulation of cancer in over 5000 patients and curate their DNA in a similar fashion to biological data. Here, we present our progress in creating and validating machine learning models, trained on our synthetic biological dataset, to more reliably identify unknown cancer genes as drivers or passengers.

Classification of Cancer Driver Genes

Cells acquire about one mutation in every 30 million base pairs during cell division processes. While most of these are harmless passenger mutations, mutations that lead to the gain or loss of cellular function drive the formation of cancer and are therefore classified as *driver mutations*. Driver genes, where driver mutations occur, are either oncogenes (activation of function) or tumor suppressor genes (loss of cancer prevention function).

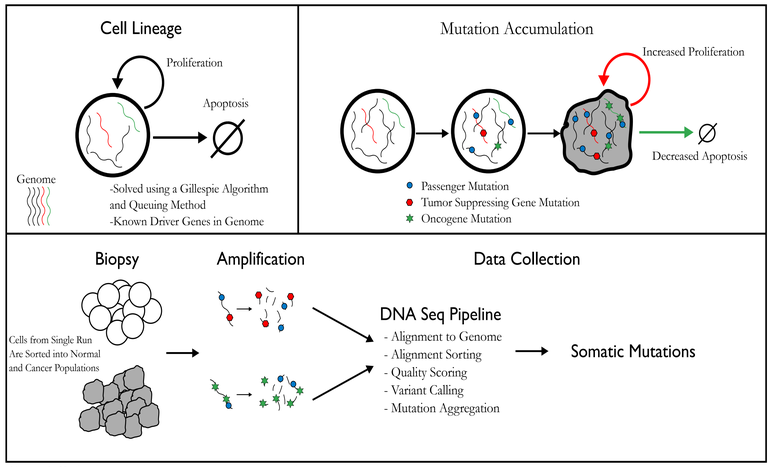


Figure 1: Definition of driver genes and the creation of our synthetic dataset.

The availability of next-generation sequencing data has allowed the development of computational classifiers of driver genes, a task relevant in precision oncology. However, these cannot be objectively assessed in precision due to the absence of a definite driver gene list. In response, we synthesized our own "gold standard" data set, by simulating the accumulation of cancer according to a mathematical model of the cancer system (see equations below). The resulting dataset contains the mutational profiles of 500,000 cancer cells, each harboring 10,000 genes with an average length of 100 base pairs.

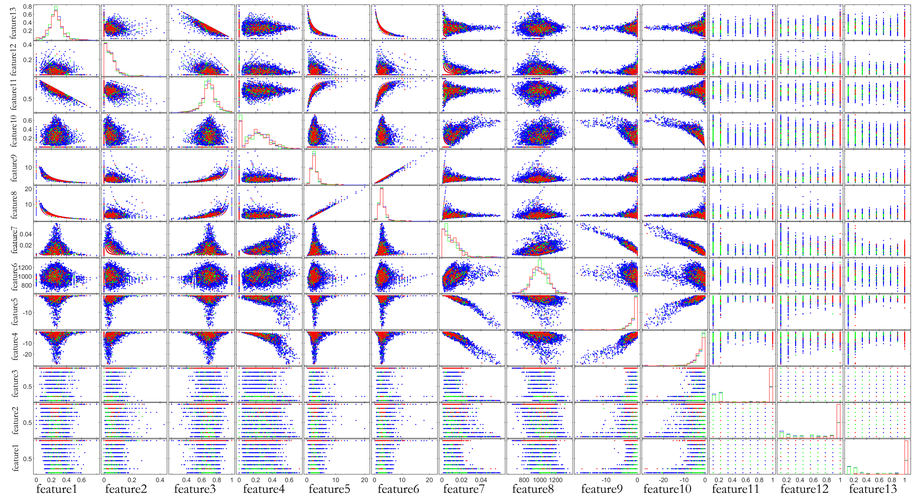
$$\frac{dN}{dt} = N(\, v_N(p_n - p_s) - d_N)(L(O_2)) - D(O_2)N$$

$$\frac{dM}{dt} = (v_M - d_M)M(L(O_2)) + v_N N(2p_s + p_a)(L(O_2)) - D(O_2)M - sIM$$

$$\frac{dI}{dt} = v_I I(1 - \frac{I}{I_H + rIM})$$

N = Normal cells, M = Cancerous (mutated) cells, I = Immune surveillance and response to cancer growth. The population of normal and cancerous cells are functions of the amount of oxygen in the system. The cancer cells gain selective advantages over time, escaping immune control and increasing their net proliferation rates (also see Fig. 1).

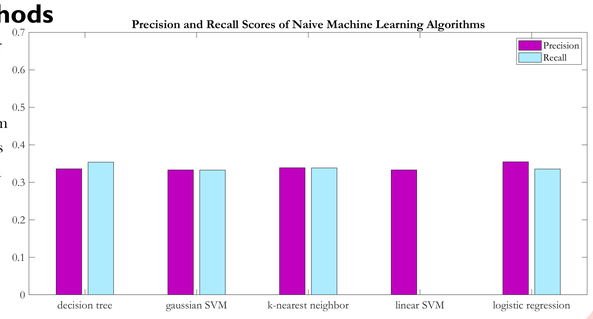
Data Exploration and Feature Selection



Thirteen features were extracted for each gene to train the classification methods with. Features 1-3 measure the frequency of each type of mutations in our synthetic data: missense, nonsense, and silent. Feature 4 is the frequency of recurrent missense mutations. Feature 5 to 6 are missense-to-silent and nonsilent-to-silent ratios. Feature 7 is the fraction of cells the mutations happen in, and feature 8 is gene length. Features 9 and 10 represent missense position entropy and nonsense position entropy. Features 11 to 13 are p-values of each mutation type's frequency, measuring the counts' significance through comparison with 10 mutation sets created through Monte Carlo simulations.

Naive Classification Methods

We first implemented and evaluated "naive", of simple, driver gene classifiers. The low precision and recall scores reflected the methods' failure to differentiate passengers from drivers. Under the assumption that our features do not lack correlation with whether a gene is a driver or passenger, we identified two characteristics of the dataset we must account for: the (1) nonlinearity of patterns and the (2) class imbalance between passengers (9.5% incidence) and drivers (2.5% OG, 2.5% TSG)



Improved Methods: Evaluation and Comparison with Existing Classifiers

We implemented (1) SVM with a Gaussian kernel and (2) random forest to handle the nonlinearly structured data. As for the problem of class imbalance, we experimented with random forest with two adaptive boosting methods that iteratively resample data points that have been misclassified in previous iterations, which in our case are likely the rarer driver genes. In addition to precision and recall, we evaluated our new models by their fraction overlap with the list of driver genes in our dataset and 'consistency', computed as the overlap between the list of top 100 drivers each predicted by classifiers trained on random halves of the data. 'Significant' genes are defined by their corresponding driver scores (probability of being a driver according to random forest) having low p-values, which are computed in comparison to the ten Monte Carlo-simulated datasets (Table 1).

The RUSBoosted random forest best classifies driver genes among the classifiers evaluated, successfully identifying 8.5% of known drivers.

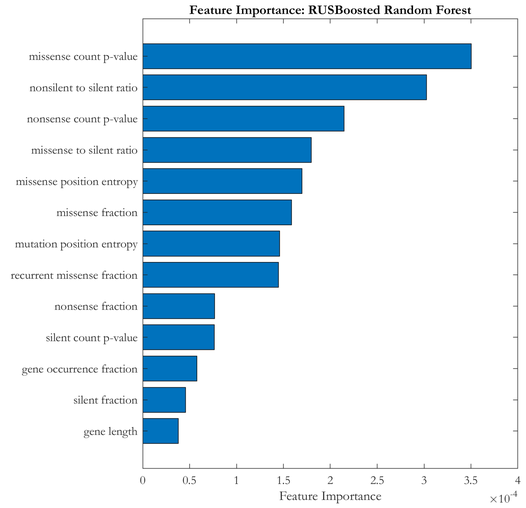
Table 1: Performance of driver gene classification methods on the synthetic biological dataset

Algorithm	# of significant genes	Fraction overlap with driver list	Precision	Recall	F1 Score	Consistency
RF: RUSBoost	956	0.85	0.6102	0.4092	0.4899	0.046
RF: AdaBoost	2088	0.75	0.4487	0.4592	0.4539	0.06
SVM: Gaussian	2036	0.208	0.613	0.3733	0.464	0.06

However, in comparison with the performance of known driver classification methods, our method's consistency is lower than those of the other classifiers (Table 2). This may imply that the drivers in our synthetic biological dataset are more similar to each other than are the drivers of the pancancer dataset. Alternatively, the low consistency may be boosted by integrating other features in the machine learning method.

Table 2: Performance of known driver gene classification methods on the pancancer somatic mutation dataset

Algorithm	# of significant genes	Fraction overlap with driver list	Consistency
2020+	208	0.4	0.749
TUSON	243	0.37	0.727
OncodriveFML	679	0.12	0.514
MutsigCV	158	0.37	0.505
OncodriveClust	586	0.07	0.232
MuSiC	1975	0.05	0.869
ActiveDriver	417	0.06	0.19
OncodriveFML	2600	0.04	0.506



The most important feature in the RUSBoosted Random Forest Algorithm is the missense count p-value; this likely contributes to the classification of oncogenes, which by definition harbor mutations on recurrent spots in the gene.

Future Work

The focus of our future work will be on improving the consistency and lowering the recall of our current random forest classifier. Based on the ranks of feature importance, the p-values, which measure the significance of some mutation frequency, are meaningful indicators of driver genes. We may experiment with other machine learning algorithms known to well handle nonlinear, imbalanced data; these include neural networks and the positive unlabeled (PU) learning algorithm. To improve the reality of the synthetic biological data, we may refine the mathematical equations that dictate the characteristics of cell growth, mutation, apoptosis, and necrosis, which then can be simulated to create a new dataset more closely resembling real-life human mutational profiles.

Acknowledgement and References

We acknowledge UCI's Center for Mathematical and Computational Biology and the MathExpLr program.

Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: Integrating quantitative models. *Nat. Rev. Cancer* **15,** 730–745 (2015).

Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci.* **107,** 18545–18550 (2010).

Vogelstein, B., Papadopoulos, N. & Velculescu, V. E. S1D_Cancer Genome Landscapes. *Science*

Whiteside, T. L. Tricks tumors use to escape from immune control. *Oral Oncol.* **45,** e119–e123 (2009).

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci.* **113,** 14330–14335 (2016).

Wilkie, K. P. & Hahnfeldt, P. Mathematical models of immune-induced cancer dormancy and the emergence of immune evasion Mathematical models of immune-induced cancer dormancy and the emergence of immune evasion. *Interface Focus* **3,** (2013).

Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173,** 371–385.e18 (2018).

Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **100,** (2000).