Spring 2017

Utilizing Tumor Exome Variation to Predict Cancer Treatment Outcomes

Michael Rendleman
University of Iowa

UTILIZING TUMOR EXOME VARIATION TO PREDICT CANCER TREATMENT OUTCOMES

by

Michael Rendleman

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Biomedical Engineering

_____

Tom Casavant
Thesis Mentor

Spring 2017

All requirements for graduation with Honors in the
Biomedical Engineering have been completed.

_____

Edwin Dove
Biomedical Engineering Honors Advisor

UTILIZING TUMOR EXOME VARIATION TO PREDICT CANCER TREATMENT OUTCOMES

by
Michael Rendleman

A thesis submitted in fulfillment of
the requirements of Engineering Honors
at the University of Iowa

May 2017
Mentor: Tom Casavant

ABSTRACT

Cancer genomics, in the context of informing clinical decisions with tumor genotype, is a field characterized by high-dimensional data. Computational approaches for evaluating sets of features to be utilized in machine learning methods are essential for yielding accurate predictive and prognostic models. Additionally, the publicly-available results of the Broad Institute's Firehose cancer genomics analysis pipeline presents a wealth of information that may be useful for cancer genotyping. Power analysis and classifier comparison are performed with the goal of evaluating a gene-based mutation significance feature set (MutSig) from Firehose. They reveal that while the MutSig features likely contain some prognostic information, the methods with which they are currently integrated do not provide enough predictive power to result in clinically-useful decision support. Results also suggest that Random Forest or other bagged classifiers are potential good candidates for feature selection and model building in this context.

INTRODUCTION AND BACKGROUND

Head and Neck Cancer

Cancer is a class of diseases characterized by abnormal cell growth and the potential for invasion of other tissues by cancerous cells. The cohort of cancer investigated in this thesis is Head and Neck Squamous Cell carcinoma (HNSC), including cancers of the oral cavity, pharynx, larynx, paranasal sinuses and nasal cavity, and salivary glands. Significant genetic alterations, or mutations, are necessary for cancer to develop. The effects of these mutations manifest as a combination of several traits: self-sufficient growth signals, insensitivity to anti-growth signals, inhibition of apoptosis, rapid reproduction, continued angiogenesis, and cell migration (or metastasis). These traits cause the formation of malignant tumors, or masses of cancerous cells.

Not all genetic alterations are created equally. Some mutations are more deleterious or cancer-causing, as different mutation locations and types can affect different biological pathways in a variety of manners. While molecular biomarkers and targeted sequencing are common in modern precision oncological medicine, the amount of genomic information that we do not know how to utilize is staggering. As a result, research broadly investigating these genotypic mechanisms is notoriously difficult, and limited success has been achieved incorporating

genotypes in clinical settings. To feasibly handle such large amounts of data, machine learning methods must be employed.

TCGA and Broad Firehose

The data utilized in this thesis was generated by The Cancer Genome Atlas (TCGA) Network, a group dedicated to producing and analyzing genomic cancer data while making their data publicly available. TCGA has released genotyping, RNA and miRNA sequencing, whole exome sequencing, methylation, and clinical data for 528 cases of HNSC. Analysis of these data has revealed many genes previously unknown to be associated with HNSC.[1]

The Eli and Edythe L. Broad Institute of MIT and Harvard (or the Broad Institute) is a well-known bioinformatics research center. The Broad Institute's Genomic Data Analysis Center (GDAC) has created a large analysis pipeline, Broad GDAC Firehose, to systematically analyze data from TCGA. The results of these analysis runs are published through online interactive figures through "Firebrowse". One tool in this pipeline, MutSig2CV (or MutSig), analyzes exomic tumor-normal variant data. For each patient in the dataset, it identifies genes that are significantly mutated above an expected baseline and reports the most deleterious disruptions to those genes (See Figure 1).[2,3]

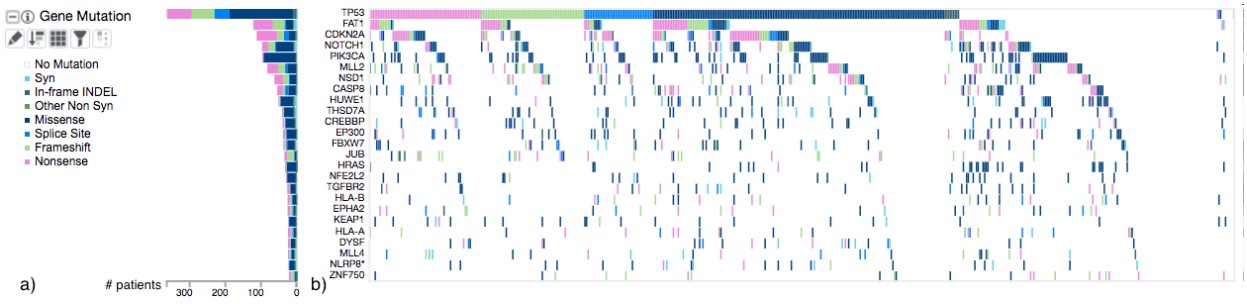

Figure 1: The output of MutSig as viewed on Firebrowse.[4] Portion (a) contains the legend and a histogram, the former listing the different types of mutations identified and the latter displaying the number of patients found with each type of mutation in each gene. Genes are vertically sorted by mutation incidence. Portion (b) displays the MutSig output itself, with each column representing a single patient in the TCGA-HNSC dataset, sorted based on TP53 mutation, followed by FAT1 mutation, then CDKN2A mutation, and so on.

Machine Learning

Machine learning (ML) is a computational field that uses algorithms to learn from existing data to discover relationships and build predictive models. Classification is a specific type of machine learning problem in which the algorithm is tasked with classifying data points

based on how previous data were classified. Essentially, data points consist of a set of features and a class label. A classifier is trained on the training data, and builds a predictive model to classify new data points.

To evaluate the effectiveness of a predictive model, it must be tested on data that it has not "seen" before. One standard, systematic way of doing this is called *k*-fold cross validation, where the available data is split into *k* partitions, or folds. Models are trained *k* times using *k*-1 of the folds as training data, while the remaining fold serves as the testing data. The performance of the classifier is then evaluated based on how it classified points in the testing data during the validation runs.

The primary goal of this research is to evaluate the Firehose MutSig2CV output in terms of how useful it might be in improving ML-based cancer outcome prediction. For this thesis, the outcome being predicted is defined as "two-year survival after diagnosis". It is worth noting that this class label results in a skewed dataset, with 69.51% of patients in the TCGA data having a positive survival outcome.

LITERATURE REVIEW

Machine learning is becoming increasingly popular in cancer research, and multiple research groups are finding moderate success in this avenue.[5,6,7] Some researchers still have reservations about some results, particularly with feature selection, as different studies do not always agree on what features are important for prediction.[6] Nonetheless, a machine learning-based approach appears to be the most promising path to deal with such complex, high-dimensional data.[8]

Guo et al. describe a general procedure for applying machine learning to high-dimensional "omics" data sets. It involves two primary steps: Dimensionality reduction and classifier training.[9] In this thesis, the selection of the Firehose MutSig data serves as dimensionality reduction, and the evaluation experiments described in Methods fall under the category of classifier training (and subsequent evaluation).

METHODS

Machine Learning Classifiers

Three types of machine learning classifiers were trained: Random Forest (RF), Multilayer Perceptron (MLP), and Glmnet.

Random Forest is a "bagged" classifier—that is, it is a classifier consisting of a group of classifiers. Essentially, 100 decision tree classifiers are trained on the training data, the trees themselves choosing random features with which to discriminate between classes. The output predicted class is decided by a majority vote of the constituent trees. Bagged classifiers are known to be more effective for complicated problems, and Random Forest has been shown to be more effective when working with skewed datasets.[9]

Multilayer Perceptron is a type of neural network, in this case having one input layer, one output layer, and $a = \frac{f+c}{2}$ hidden layers, where f is the number of unique features in the training set, and c is the number of classes. Each node in the network uses a sigmoid kernel function.

Glmnet is an algorithm that fits a generalized linear model via penalized maximum likelihood. In this case, the model built is a type of logistic regression model. It also utilizes lasso regularization, a regression analysis method that performs both regularization and feature selection to improve predictive accuracy.[10,11]

The Random Forest and Multilayer Perceptron classifiers were trained and tested using their respective implementations in Weka 3.9.1, and Glmnet models were trained in Weka with the Glmnet R package through the RWeka interface. All classifiers used the default settings from their implementation and 10-fold cross validation unless otherwise noted.

Classifier Comparison Metrics

Three metrics were used to compare classifiers: accuracy, the area under the Receiver Operating Characteristic curve (AUROC), and the area under the Precision-Recall curve (AUPRC). While high accuracy is certainly desirable, it makes for a poor metric of predictive power. The Receiver Operating Characteristic (ROC) curve is a plot of true-positive rate versus false-positive rate while varying a discriminative threshold within the classifier. The area under this curve (AUROC, often denoted AUC) is a standard measure for predictive power in machine learning literature. The Precision-Recall curve is a plot of Precision versus Recall, and the area under this curve (AUPRC) is also a valuable measure of predictive power. Additionally, the AUPRC metric has been shown to be more discriminative in skewed datasets than AUROC

while being no less informative.[12] Since this classification problem has prominent class imbalance, AUPRC was the primary consideration in model evaluation, followed by AUROC.

Clinical-MutSig Evaluation

Classifiers were trained on 43 clinical data features, and then trained on a combination of the clinical data and 10 gene-based mutation significance (MutSig) features. The full list of clinical and mutation significance features used in this thesis can be found in Supplementary Tables 1 and 2, respectively. The description for Supplementary Table 2 also describes some feature selection and preprocessing of the mutation significance features.

Clinical-MutSig Power Evaluation

To examine the mutation significance data more directly, the mutation significance features were permuted by shuffling values randomly within each feature. These shuffled mutation significance features were then used with the actual non-shuffled clinical features to train the classifiers. A total of ten shuffled mutation significance datasets were created and used in this way. The mean and standard deviation of the predictive power metrics for these classifiers were calculated, and one-sided t-tests were performed with the null hypothesis that the classifiers trained with the actual mutation significance data were not more predictive than the classifiers trained with randomly-shuffled mutation significance data.

T-values were calculated with $t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, where $\bar{x}$ is the metric under consideration, calculated for the classifiers trained with actual mutation significance data, $\mu_0$is the mean of the metrics calculated for the classifiers trained with shuffled mutation significance data, σ is the standard deviation of the shuffled metrics, and *n* is the number of samples (in this case 10). P-values were obtained for the one-sided significance tests with the Microsoft Excel function TDIST.

This permutation-based evaluation model is similar to the recursive feature selection algorithm utilized by Guo et. al for performing analysis of high-dimensional data.[9]

MutSig Power Evaluation

The randomly-shuffled mutation significance data from the previous power evaluation method were utilized for classifier training, but without the presence of the more predictive clinical data. Additionally, a second set of shuffled data was created, except the class labels for the examples were permuted instead of the mutation significance features. As before, classifiers

were trained on both the actual and permuted datasets, and statistical t-tests were performed similarly to the Clinical-MutSig power evaluation step. For this step, Glmnet models were not trained, as automated training and evaluation of this classifier for multiple datasets proved difficult due to the nature of the RWeka package. In addition, the Glmnet models had not proved especially effective at extracting predictive power from these data in previous evaluation steps.

RESULTS

Result 1: Clinical-MutSig Evaluation

	MLP			RF			Glmnet		
	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC
Clinical	69.89%	0.717	0.745	70.45%	0.696	0.721	68.37%	0.687	0.711
w/ MutSig	71.21%	0.717	0.738	69.89%	0.704	0.734	68.37%	0.687	0.712

Table 1: Classifier performance with both Clinical and Clinical + MutSig feature sets. Baselines for each metric are as follows: Accuracy: 69.51%. AUROC: 0.500. AUPRC: 0.695.

For the MLP and Glmnet classifiers, either no change or a decrease in predictive power metrics was observed with the addition of mutation significance data. With the RF classifier, a small increase in AUPRC was observed, though the statistical significance of this increase could not be confirmed.

Result 2: Clinical-MutSig Power Evaluation

	MLP			RF			Glmnet		
	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC
p-value	< 5e-4	< 0.001	< 0.1	< 0.2	< 0.005	< 5e-4	> 0.2	< 0.025	< 0.1

Table 2: Results of one-sided t-tests for each comparison metric, as described in "*Clinical-MutSig Power Evaluation*" in Methods. Highlighted entries are cases where $p < 0.05$.

Significant differences were found in AUROC for RF, MLP, and Glmnet classifiers between the actual and shuffled data, with the actual mutation significance data performing better. Additionally, the real data showed significantly better performance in AUPRC for the RF classifier only.

Result 3: MutSig Power Evaluation

	MLP			RF		
	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC
p (gene shuffle)	< 2.5e-6	< 1e-5	< 5e-5	< 1e-4	< 2.5e-7	< 2.5e-6
p (class shuffle)	< 2.5e-6	< 2.5e-5	< 2.5e-5	< 5e-4	< 1e-4	< 2.5e-4

Table 3: Results of one-sided t-test for each comparison metric, as described in "*MutSig Power Evaluation*" in Methods. Highlighted entries are cases where $p < 0.05$.

Significant differences were found in accuracy, AUPRC, and AUROC for both RF and MLP (Glmnet models were not trained for this evaluation). In both the case of within-feature shuffling and the case of class label shuffling, the actual mutation significance data outperformed the shuffled data much more clearly than in the previous evaluation which included clinical data features.

DISCUSSION

The results in Table 1 show little to no improvement with the simple addition of the mutation significance features to the dataset. Glmnet, a method that automatically selects the most informative and non-correlative features, shows no change between the clinical and mutation significance-enriched datasets. This indicates that the clinical data may provide the same or much more predictive information than the MutSig data. To reduce the influence of the clinical features on the analysis, the power evaluations were performed.

In Table 2, results show that the mutation significance features are not devoid of predictive power, and the p-values for AUPRC and AUROC of the Random Forest classifier confirm the results of Guo et. al[9] regarding the relative effectiveness of this classifier with class imbalance problems. This suggests that RF and/or other bagged classification methods may be good candidates for model building in this context.

The results of the third evaluation show even more clearly that predictive value exists in the mutation significance data. These results also confirm the conclusions reached about the relative information present in the clinical data compared to the mutation significance data.

It is worth noting that in many cases, the MLP classifier tended to drastically overfit the data, often yielding training accuracies 20 percentage points higher than the testing accuracies. This would indicate that metrics obtained from MLP classifiers in this thesis may be misleading, and emphasizes the importance of parameter tuning during model training. The RF and Glmnet classifiers showed no such evidence of overfitting.

CONCLUSION

Overall, it appears that the MutSig data as output by Broad's Firehose analysis pipeline does have some predictive value when considering two-year survival of HNSC patients in TCGA. Results indicate that the information provided by these mutation significance data is either correlative with, or dwarfed by the information found in clinical features. Additionally,

results with the Random Forest classifier suggest that it or other bagged classifiers may be effective tools for this classification problem.

Though some predictive power has been found in the data, this approach is not one that can be directly extended to clinical environments. However, the presence of any predictive information suggests that further high-level or even variant-level analyses may prove effective in increasing predictive power.

ACKNOWLEDGEMENTS

Tom Casavant Brian Smith

John Buatti Bartley Brown

Terry Braun Jonathon Tessmann

Katelyn Welander Jon Kuhl

Hanson Center for Technical Communication

REFERENCES

[1]The Cancer Genome Atlas, N. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature,* 517**,** 576-582.

[2]Broad Institute Cancer Genome Analysis. (2013). MutSig. Retrieved from http://archive.broadinstitute.org/cancer/cga/mutsig

[3]Broad Institute TCGA Genome Data Analysis Center. (2017, February 3). Firehose Documentation. Retrieved April 26, 2017, from https://confluence.broadinstitute.org/display/GDAC/Documentation

[4]Broad Institute TCGA Genome Data Analysis Center (2016): Firehose stddata__2016_01_28 run. Broad Institute of MIT and Harvard. doi:10.7908/C11G0KM9

[5]Aerts, H. J. W. L., et al. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications,* 5**,** 4006.

[6]Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis & D. I. Fotiadis (2015) Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal,* 13**,** 8-17.

[7]Kang, J., R. Schwartz, J. Flickinger & S. Beriwal Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective. *International Journal of Radiation Oncology • Biology • Physics,* 93**,** 1127-1135.

[8]Larranaga, P., et al. (2005). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7, 86-111.

[9]Guo, Y., A. Graber, R. N. McBurney & R. Balasubramanian (2010) Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *BMC Bioinformatics,* 11**,** 447.

[10]Hastie, T., & Qian, J. (2014, June 26). Glmnet Vignette. Retrieved April 26, 2017, from https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

[11]Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society*. Series B (methodological) 58 (1). Wiley: 267–88. http://www.jstor.org/stable/2346178.

[12]Davis, Jesse, & Goadrich, Mark. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning (ICML '06)*. ACM, New York, NY, USA, 233-240. doi: 10.1145/1143844.1143874

SUPPLEMENTARY MATERIALS

Supplementary Table 1: List of clinical data features, including patient- and treatment-specific information.

Age	Alcohol consumption per day	Anatomic Organ
Ethnicity	Laterality	Lymphovascular Invasion
Margin Status	Tobacco Pack-Years Smoked	Perineural Invasion
Lymph Nodes Examined	HPV Status (ISH)	HPV Status (P16)
Lymph Node Neck Dissection Indicator	Extracapsular Spread, Pathologic	Smokeless Tobacco Average per day
Gender (sex)	Race	Tumor Grade
Tobacco Smoking History	History of Neoadjuvant Tx	AJCC Clinical Nodes (CN)
AJCC Clinical Metastasis (CM)	AJCC Pathologic Metastasis (PM)	AJCC Pathologic Tumor Stage
AJCC Pathologic Tumor (PT)	AJCC Clinical Tumor (CT)	AJCC Clinical Tumor Stage
Radiation Tx Adjuvant	AJCC Pathologic Nodes (PN)	Pharmaceutical Tx Adjuvant
New Tumor Event Pharm Tx	Pharmaceutical Therapy Adj.	New Tumor Event Surgery
New Tumor: Surgery Metastatic	New Tumor: Surgery Locoregional	Adjuvant Radiation Fractions Total
Radiation Total Dose	New Tumor Event Rad Tx	Definitive Tx Method
RX: Therapy Type	Radiation Therapy Type	RX: Number of Cycles
RX: Total Dose		

Supplementary Table 2: List of genes for which mutation significance was considered. The ten most-populated genes were chosen from the MutSig2CV output. All gene features other than TP53 were considered too sparsely-populated for informative categorical distinction (77%-93% with "no mutation"), and as such were collapsed into binary variables ("mutation" or "no mutation") to prevent overfitting. The TP53 mutation significance feature itself exhibits a distribution within which individual categories are relatively well-populated, with only 32% of examples having "no mutation".

TP53	FAT1
CDKN2A	NOTCH1
PIK3CA	MLL2
NSD1	CASP8
HUWE1	THSD7A